Discussion Papers of the
Max Planck Institute for
Research on Collective Goods
2022/4

Fair Governance with Humans and Machines

Yoan Hermstrüwer
Pascal Langenbach

MAX PLANCK
SOCIETY

Discussion Papers of the
Max Planck Institute
for Research on Collective Goods

2022/4

Fair Governance with Humans and Machines

Yoan Hermstrüwer / Pascal Langenbach

May 2022

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
https://www.coll.mpg.de

Fair Governance with Humans and Machines[*]

This version: 23 May 2022

Yoan Hermstrüwer[†] Pascal Langenbach[‡]

How fair are government decisions based on algorithmic predictions? And to what extent can the government delegate decisions to machines without sacrificing procedural fairness? Using a set of vignettes in the context of predictive policing, school admissions, and refugee-matching, we explore how different degrees of human-machine interaction affect fairness perceptions and procedural preferences. We implement four treatments varying the extent of responsibility delegation to the machine and the degree of human involvement in the decision-making process, ranging from full human discretion, machine-based predictions with high human involvement, machine-based predictions with low human involvement, and fully machine-based decisions. We find that machine-based predictions with high human involvement yield the highest and fully machine-based decisions the lowest fairness scores. Different accuracy assessments can partly explain these differences. Fairness scores follow a similar pattern across contexts, with a negative level effect and lower fairness perceptions of human decisions in the context of predictive policing. Our results shed light on the behavioral foundations of several legal human-in-the-loop rules.

Keywords: algorithms; predictive policing; school admissions; refugee-matching; fairness.

[*] We would like to thank Christoph Engel, Nina Grgić-Hlača, Laurence O'Hara, and participants of the MD4SG Working Group on Algorithms, Law, and Policy (Spring 2021), of the 2021 Conference of the International Society of Public Law (ICON-S Mundo), of the 18th Annual Conference of the German Law and Economics Association (GLEA 2021), and of the 38th Annual Conference of the European Association of Law & Economics (EALE 2021). We are grateful to Brian Cooper for carefully proofreading the latest version of this article.

[†] E-mail: hermstruewer@coll.mpg.de, Address: Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113, Bonn, Germany.

[‡] E-mail: langenbach@coll.mpg.de, Address: Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113, Bonn, Germany.

1. Introduction

Machine-learning algorithms are increasingly used to predict risks and assist public officials in their decisions. While the initial discussion has focused on the algorithmic assistance of judges in sentencing, pre-trial, or bail decisions (Kleinberg et al., 2017), similar algorithmic decision aids are rapidly expanding to other areas of public decision-making (Huq, 2020b). Some of the most notable applications include the prediction of crime in order to deploy police forces more effectively (Joh, 2016; Simmons, 2018), the matching of refugees with municipalities based on predicted integration success (Acharya, Bansak, & Hainmueller, 2022; Ahani et al., 2021; Bansak et al., 2018), and the admission of students to schools based on their chances of completing their degree (Kearns & Roth, 2019; Muratov et al., 2017).

In this article, we address the perceived fairness of such algorithmically assisted decision procedures in the public sector. In particular, we are interested in how procedural fairness perceptions vary with the degree of machine involvement in the decision-making process, and whether fairness perceptions systematically differ across different legal or policy contexts.

With the increasing application of machine-learning algorithms, the perceived fairness of algorithmic decision-making becomes increasingly important. This holds especially for algorithmic decision aids in public-sector decision-making. The perceived fairness of algorithmically assisted procedures is an important yet underrated precondition for the success of algorithmic governance, for procedural justice is one of the cornerstones of legal compliance and of the legitimacy of government actions (Wang, 2018; Tyler, 2006, 2003).

Moral motivations and the desire to be treated fairly are important forces underlying legal compliance, yet these motivations are not fully captured by narrower versions of rational choice theory that usually consider material self-interest to be the main source of legal compliance (see, e.g., Becker, 1968). The fairness perceptions of those potentially affected by algorithmic predictions or classifications might therefore be predictive for the future role of algorithms in public-sector decision-making (Nagtegaal, 2021).

To explore our research questions, we conduct an online experiment based on a set of vignettes covering three areas of public-sector decisions: predictive policing, school admissions, and refugee-matching. Treatments differ in whether (i) a human, (ii) an algorithm, or (iii) a human assisted by an algorithm makes the decision. The latter case is split into two treatments: one in which the algorithm's assessment of the facts only provides additional information for the human decision-maker, and one in which the human more often than not just relies on the algorithm's assessment, hence practically delegating the decision to the machine in most of the cases. We measure the perceived fairness of these procedures in a sample of participants recruited from Amazon Mechanical Turk (MTurk). Without knowing the outcomes of these procedures for any particular case, participants judge the fairness of the procedure they are presented with.[1]

Our results indicate that algorithmically assisted decision procedures with a high degree of human involvement yield the highest procedural fairness scores. Fully human decision procedures and algorithmically assisted decision procedures with a low degree of human control are evaluated as equally fair. Fully algorithmic decision procedures, by contrast, fare worst in terms of procedural fairness. Overall, this suggests a prevalence of strong fairness preferences for hybrid decision

[1] The treatment variations we implement in this study are also related to earlier work in experimental economics, especially research on ultimatum bargaining (Güth, Schmittberger, & Schwarze, 1982; Roth, 1995). Several studies show that respondents accept lower offers when the split is determined by a computer (Inaba et al., 2018; Blount, 1995), they are more likely to reciprocate a helpful offer in case of a human offer (Offerman, 2002), and, more generally, they show stronger reciprocal responses when confronted with a human offer (Falk, Fehr, & Fischbacher, 2008).

procedures with a high degree of human involvement. However, people do not seem to care much whether a *human does all* the work or whether a *machine does most* of the work. These results provide important guidance for the interpretation and for the design of legal rules aimed at organizing the division of labor between humans and algorithms.

The remainder of this article proceeds as follows. In the next section, we discuss how our study contributes to the literature on algorithmic public decision-making. Section 3 presents our research design. In Section 4, we report our results and discuss their relevance for the literature. Section 5 concludes.

2. Literature

Our study contributes to the literature on algorithmic fairness in public decision-making on three levels. First, we add to a newer strand of research that explicitly focuses on the interaction between algorithms and human decision-makers (Imai et al., 2021; Green & Chen, 2019). Considering the risks of discrimination, in-group bias, or automation bias in algorithmic decision-making, legal scholars have been discussing whether and to what extent the law actually grants a right to a human decision (Huq, 2020a). Computer scientists have also voiced claims in favor of human-in-the-loop, human-on-the-loop, or human-in-command requirements (Binns, 2020; Yaghini, Heidari, & Krause, 2021). This corresponds to the basic model of Art. 22 (1) EU General Data Protection Regulation, formulating the principle that no person shall be subject to a decision based on fully automated data processing.[2] Under Art. 14 (1) of the proposed EU Artificial Intelligence Act (AI

[2] Art. 22 (3) GDPR contains several exceptions to this principle. This indicates that the material scope of the right to a human decision may be context-dependent rather than universal, as several use cases will likely be exempted from the right.

Act), high-risk AI systems, for example predictive schooling systems like the one we explore in this study, shall be designed and developed in such a way that they can be effectively overseen and fully understood by humans. Others have been more optimistic about the future of purely machine-made decisions and have argued that the outputs generated by machine-learning algorithms should be used as micro-directives (Casey & Niblett, 2017). Current algorithmic decision-making practices, however, are based on the premise that decisions cannot or should not be entirely delegated to a machine. Rather, they are based on some interaction between a human decision-maker and an algorithmic decision aid. While the recent literature has included hybrid decisions as a third category in the spectrum spanning fully human and fully algorithmic decisions (Nagtegaal, 2021), only little attention has been paid to the effects of different degrees of control in the interaction between humans and algorithms. Therefore, in addition to comparing the perceived fairness of human and algorithmic decision procedures, we also explore procedures in which human decision-makers are assisted by algorithmic decision aids and exert different levels of control over the final outcome.

Second, with our set of vignettes covering predictive policing, school admissions, and refugee-matching, we can compare the fairness of the different algorithmic decision aids in three practically relevant public-law contexts. Studies in several academic fields have assessed the perceived fairness of algorithms in governmental and legal contexts (for a recent review of the empirical literature, see Starke et al., 2021), but the overwhelming majority of these studies focus on the criminal-justice system (see, e.g., Imai et al., 2021; Harrison et al., 2020; Dodge et al., 2019; Grgić-Hlača et al., 2018a, 2018b; Simmons, 2018; Wang, 2018). While algorithmically assisted decision-making has indeed been very prominent in the context of criminal justice, it is difficult to extrapolate results from the criminal justice context to other contexts. Only a few studies have extended the relatively narrow contextual scope of existing studies, exploring fairness perceptions in the context of university admissions (Marcinkowski et al., 2020), parking offenses and criminal charges (Araujo et al., 2020), child protective services and unemployment aid (Albach & Wright, 2021),[3] and the enforcement of traffic laws (Miller & Keiser, 2021). Our study is designed to generate evidence that is more robust across different areas of the law, thus exploring legal decision-making procedures beyond the criminal-justice context.

Third, we study the *perceived* fairness of algorithmic decision procedures. Algorithmic fairness can be conceptualized in different ways. One line of research studies the fairness of algorithmic predictions from an objective or normative perspective. This research ultimately tries to improve algorithmic predictions measured by some normative standard, such as statistical parity, equality of false-positives, equality of false-negatives, or equality of predictive accuracy (see, e.g., Barocas, Hardt, & Narayanan, 2021; Berk et al., 2021; Hellman, 2020; Kleinberg, Mullainathan, & Raghavan, 2017; Chouldechova, 2017; Corbett-Davies et al., 2017). In the tradition of fairness research in the social sciences, another line of research takes a subjective approach and is concerned with the perceived fairness of algorithmic decisions among potential addressees or in the public. A common distinction is made according to the object of fairness judgments, i.e., whether they refer to decision outcomes (*distributive fairness*) or to the decision-making process (*procedural fairness*) (see Lind & Tyler, 1988; Walker, Lind, & Thibaut, 1979). While achieving distributive fairness may be an important element of legitimacy, for example by defining a social-welfare function that captures a preference for more equitable outcomes (Rambachan et al., 2020), the guarantees of procedural fairness are no less important in legal terms (see Tyler, 2006).

[3] Albach and Wright (2021) additionally investigate the fairness of specific features in the context of bail, hospital resources, insurance rates, and loans.

If fairness perceptions matter for legal compliance, and if the legal order is keen on achieving effectiveness and legitimacy, legal scholars and policy-makers cannot simply dodge the question how much human involvement exactly the law should guarantee in algorithmically assisted decision-making. Recognizing the behavioral dimension of fairness, a growing literature especially in computer science has turned its attention to fairness perceptions of algorithmic decision procedures (for a summary, see Starke et al., 2021). One key insight of this literature is that fairness perceptions seem to be highly context-dependent (Starke et al., 2021). This suggests that it may be difficult to derive general conclusions about the relative fairness of algorithmic and human decisions. This is mostly due to the lack of consistent behavioral patterns uncovered in existing empirical studies.

On the one hand, empirical evidence suggests a fairness preference for human decision-making processes.[4] Chen, Stremitzer, & Tobia (2022) report evidence from a vignette study – with three scenarios covering a consumer refund, a pre-trial bail decision, and a custodial sentencing decision – and show that a human judge is perceived as fairer than an algorithmic judge. Focusing on decisions in the criminal-justice context, and using a representative sample of the US population, Wang (2018) reports in several vignette studies that the use of a computer algorithm in bail decisions is disliked compared to other expert decision procedures, with fairness perceptions being affected by information about the accuracy of the procedure. Yet, people's dislike for algorithms in bail decisions depends not just on the accuracy of such decisions, but also on the distribution of false-positive rates across groups (Harrison et al., 2020). However, within a sample of 600

[4] This strand of literature is in line with more general evidence showing that people prefer human over algorithmic decisions (see, e.g., Lee & Baykal, 2017; Lee et al., 2019) and that humans tend to distrust algorithmic outputs, a phenomenon sometimes referred to as *algorithm aversion* (Dietvorst, Simmons, & Massey, 2015).

participants, Simmons (2018) reports no differences in fairness perceptions between bail decisions made by a judge with or without the assistance of a "computer program".

On the other hand, studies also show that people assess automated decision-making as fairer than the human alternative.[5] Araujo et al. (2020), for example, report similar fairness perceptions of algorithmic and human decisions across different contexts. However, when the consequences of decisions are severe, people judge algorithmic decision-making as fairer (for example, the administrative decision whether to issue a fine for wrong parking vs. the prosecutorial decision to bring criminal charges). In an experiment on policing by Miller & Keiser (2021), black participants prefer traffic control by automated red-light cameras to a police officer when shown a picture that suggests an underrepresentation of black citizens in the municipal police department. In a survey study, Marcinkowski et al. (2020) find that students rate university admissions decisions made by an algorithm as fairer, and the decision procedure as less biased, compared to a human admissions committee. Studying fairness perceptions of public employees, Nagtegaal (2021) finds that human decision-making is perceived as fairer than fully algorithmic decision-making for more complex tasks that cannot easily be quantified, whereas the ranking was the other way around for simpler tasks. Descriptively, a combination of a human and an algorithm was in the middle but not statistically different from human decision-making.

In light of these inconclusive results, further empirical investigations of the procedural fairness of algorithmic legal decision-making are inherently valuable. A broader contextual scope including a diverse set of policy areas and decision procedures is likely to contribute to a better understanding

[5] This strand of literature is in line with evidence showing that humans tend to appreciate the use of algorithms in specific commercial contexts, a phenomenon sometimes dubbed *algorithm appreciation* (Logg, Minson, & Moore, 2019).

of what drives the acceptance of algorithmic decisions and how far the delegation of responsibility can go without sacrificing procedural fairness.

3. Research Design

3.1. Procedures

To explore our research question, we conducted a vignette study on MTurk. The study was programmed in Qualtrics and deployed through CloudResearch to ensure a reliable recruitment of participants.

Our sample consists of 1598 participants, recruited from the MTurk marketplace in the US, as all algorithmic decision support systems we explore in our study have been either developed or predominantly applied in the US to this date. More than 50% of participants are *younger than 35 years* and approximately 10% are *older than 54 years*. With 63%, men are over-represented in our sample. 67% of our sample identify as *White*, 26% as *Black or African American*, and 5% as *Asian*.[6] Roughly, 60% report a *four-year college degree* as their highest education, and over 18% report a *professional degree*.

Aware of the challenges posed by MTurk (see Horton, Rand, & Zeckhauser, 2011), we implemented a variety of measures to enhance the validity of our results. To mitigate further potential self-selection problems, we ran the study in different sessions on different days and at different times of day to ensure a diverse composition of the participant pool. To motivate

[6] Our socio-demographic sample composition seems relatively close to the numbers reported by the United States Census Bureau (as of 1 July 2021, < https://www.census.gov/quickfacts/fact/table/US/PST045221>), although we expect that several Hispanics or Latinos identified as either White or Black or African American in our racial survey classification.

participants to engage seriously with the vignettes, we made sure to keep our vignettes short and paid a competitive participation fee. On average, participants spent approximately 8 minutes on the vignettes and earned 1.50 USD after completing the study. In addition, we implemented an attention check before participants began reading the vignettes. Only participants who passed the attention check entered our sample. We also imposed a time constraint of 45 minutes to ensure that participants devoted their full attention to the vignettes.

3.2. Treatments

In a between-subjects design, we study four treatments that differ in the extent to which the decision is based on algorithmic assistance. This design choice is motivated by the observation that algorithmically assisted decision procedures vary in the level of automation (see Cummings, 2017; Manzey, Reichenbach, & Onnasch, 2012). The midpoint of the spectrum between fully human and fully algorithmic decision procedures separates executions that the human needs to *approve* and procedures allowing humans to *veto* an otherwise automatic execution.

In the *HUMAN* treatment, the decision is entirely made by a human decision-maker and solely based on a human assessment of the facts. Participants therefore read that a human decision-maker will *conduct an in-depth analysis of the case material* and *assess the risk* or the *success probability*. Participants also read that the human decision-maker *has discretion* in making the decision. On the other end of the spectrum, in the *MACHINE* treatment, the decision is entirely controlled by a computer algorithm. Participants read that a computer algorithm will conduct the *in-depth analysis of the case material* and *assess the risk* or the *success probability*. Further, the computer algorithm will make the final decision that no human decision maker can override.

Between those extremes, we implement two treatments with algorithmically assisted decision-making. In both treatments, a human who has discretion in making the decision takes the final decision. Yet the degree of algorithmic assistance and the level of human involvement and control – high or low – differs between treatments. In the *HIGH* treatment, the computer's assessment of the facts and the resulting probabilities are *always* accompanied by a human assessment. Participants therefore read that *the decision will never be based on the computer algorithm alone*, but that the human decision-maker *will always conduct his or her own analysis* before making the final decision. In the *LOW* treatment, by contrast, the human input in the decision-making process is heavily reduced as *the decision will usually be based on the computer algorithm alone*. The human decision-maker will only *sometimes conduct his or her own analysis*, meaning that the human decision-maker *will in some cases conduct an in-depth analysis of the case material, and assess the risk/success probabilities*.

The descriptions of the computer algorithm and of the human assessment are identical across all treatments (when applicable). While our vignettes contain a precise description of the facts that the computer algorithm and the human decision-maker use to make their assessments and how these facts are elicited, by design, we keep the mechanics of the computer algorithm vague. Given that we are interested in the fairness evaluations of lay people, we deem it externally valid to give no further information about the technical details of the algorithm, since the public will most likely not have more detailed knowledge about how a computer algorithm assisting a government official produces its results.

3.3. Scenarios

Our main research interest focuses on fairness perceptions of different forms of algorithmic assistance in public-sector decision-making. We explore these differences based on between-subject treatment comparisons. However, in order to enhance the robustness of our findings across different practically relevant areas of the law, we implement each of the four treatments in three different scenarios. In this within-subjects component of our experiment, participants in a session respond to one treatment presented in three different legal contexts: a predictive-policing scenario, a school-admissions scenario, and a refugee-matching scenario (Figure 1).

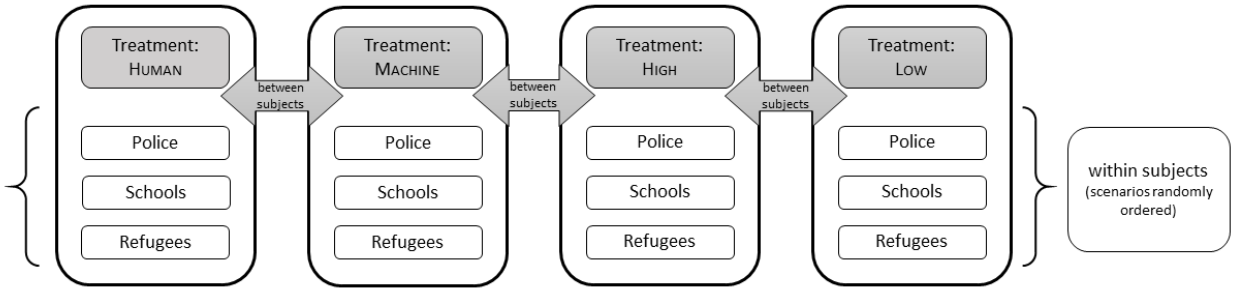


Figure 1: Structure of the experiment

Hence, for a given treatment, each participant reads all three scenarios. Scenarios are presented in randomized order. Example vignettes for the different treatments and scenarios are shown in Figure 2.

Apart from representing different policy contexts, the three scenarios also differ in other regards. First, the task of the computer algorithm and the goal of the human assessment slightly differ across the different scenarios. In the predictive-policing scenario, it is the risk of violent crimes in specific areas of the city that needs to be predicted. In the school-admissions scenario, the probability of graduation is assessed, whereas in the refugee scenario the probability of employment for a refugee in a certain location is of interest. Second, in the predictive-policing and the refugee scenario, a single human decision-maker, either a police or an immigration officer, is in charge. In the school scenario, a school admissions board manages the application procedures and decisions. Third, while the tasks used in all our vignettes are not purely mechanical and easily quantifiable, they slightly differ in the level of complexity. Predicting crime in a certain area may be simpler than predicting the probability of employment of refugees, as the latter is likely to depend on individual characteristics as well as fluctuations in supply and demand in labor markets. Predicting the employment of refugees may in turn be simpler than predicting success at school, as this depends on individual characteristics and the evolution of skills over a long period of time. Task complexity might also affect the relative evaluation of human or algorithmic decision procedures (Nagtegaal, 2021).

Treatment: Human - Scenario: *Refugees*	**Treatment: Machine - Scenario: *Schools***	**Treatment: High - Scenario: *Police***	**Treatment: Low - Scenario: *Police***
One of the main tasks of immigration authorities is to assign refugees to certain locations within the country of immigration. Refugee facilities have limited capacities. Therefore, the immigration authorities have to assign refugees based on some criterion. One prominent criterion is the chance that a refugee will be able to integrate herself into society. In applying this criterion, immigration authorities usually assess the probability that the refugee will successfully find employment when assigned to a certain location.	Many public schools have limited capacities. Accordingly, these schools are unable to accept all students who apply. Therefore, the school admissions boards have to select students based on some criterion. One prominent criterion is the chance that an applicant will succeed within the respective school system. In applying this criterion, school admissions boards usually assess the probability that the applicant will eventually graduate.	One of the main tasks of the police is to prevent criminal behavior. In order to deploy their forces in an optimal manner, the police need to assess the risk that criminal behavior will occur. This risk assessment refers to various types of criminal behavior, including the risk of violent assaults.	One of the main tasks of the police is to prevent criminal behavior. In order to deploy their forces in an optimal manner, the police need to assess the risk that criminal behavior will occur. This risk assessment refers to various types of criminal behavior, including the risk of violent assaults.
Suppose an immigration authority wants to assess this probability and decide to which location within the country of immigration a refugee should be assigned	Suppose a school admissions board wants to assess this success probability and decide whether to accept or reject an applicant.	Suppose the local police want to assess the risk of violent assaults in certain areas of the city - including the probable type, location, and time of the assault - and perform bodily searches of all persons within a small and well-defined area of the city. The purpose of these bodily searches is to track down weapons used for violent assaults.	Suppose the local police want to assess the risk of violent assaults in certain areas of the city - including the probable type, location, and time of the assault - and perform bodily searches of all persons within a small and well-defined area of the city. The purpose of these bodily searches is to track down weapons used for violent assaults.
The case manager will conduct an interview with the refugee and collect information on the refugee’s origin, education, linguistic skills, and family status. Then the case manager will conduct an in-depth analysis of the case material and assess the probability of successful employment.	The school admissions board will conduct an interview with the applicant and collect additional information on the applicant’s school history, extracurricular engagement, the distance between the school and the home, and family support. Then the admissions board will use a computer algorithm to assess the applicant’s success probability. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the applicant’s success probability to the admissions board.	The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer.	The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer.
	The decision **will be based** on the computer algorithm's assessment **alone**.	The decision **will never be based** on the computer algorithm **alone**. The police officer will **always** conduct his or her own analysis, that means, the police officer will in each case conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.	The decision will **usually be based** on the computer algorithm **alone**. The police officer will **sometimes** conduct his or her own analysis, that means, the police officer will in some cases conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.
Based on his or her assessment of the probability of successful employment, the case manager will assign the refugee to a certain location. The case manager has discretion in this decision.	Based on its assessment of the success probability, the computer algorithm will accept or reject the applicant. The admissions board cannot override the decision of the computer algorithm and has no discretion in this decision.	Based on the risk assessment of the computer algorithm and his or her own risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.	Based on the risk assessment of the computer algorithm and - only if conducted - his or her own risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.

Figure 2: Vignette examples for different treatments and scenarios

3.4. Dependent Variables

After reading the vignette for each scenario, participants answered four questions. First, we asked participants to indicate the fairness of the procedure by which the decision-maker would come to her decision. Participants could choose one of seven possible answers ranging from *very unfair* (1) to *very fair* (7). Second, as a control variable, we asked participants for their accuracy assessment of the probability estimate on which the decision-maker would base her decision. Different evaluations of the accuracy of a procedure might help explain differences in fairness ratings (Wang, 2018). Participants could choose one of seven possible answers, ranging from *not accurate at all* (1) to *extremely accurate* (7).

Additionally, we elicited responses to two questions designed to identify whether the participants' fairness assessments differ between situations in which they are personally involved or not involved (see Wang, 2018). Therefore, we asked participants whether they would want the decision-making procedure to be implemented in case they were personally affected by the decision. Finally, we asked whether they would want the procedure to be used for the public. In both cases, participants could choose one of seven possible answers, ranging from *not at all* (1) to *to a large extent* (7).

To control for socio-demographic characteristics, after the last vignette, we also collected individual-level covariates, including age, education, gender, ethnicity, political affiliation, and the weekly hours spent on remunerated tasks like those offered on MTurk.

3.5. Hypotheses

We designed our study to test the effects of different degrees of human control in different contexts of public-sector decision-making on procedural fairness. The previous literature has found preferences for human decision-making as well as machine-based decision-making in different contexts. Therefore, we generally expect that the degree of human (machine) involvement in the decision-making procedure affects the participants' fairness perceptions:

H1: Fairness ratings react to the different degrees of human involvement in the different treatments.

Many studies in the realm of legal decision-making report a preference for human decision-making. If the participants show this preference in our study, too, we expect the fairness ratings to increase with human involvement in the decision-making process. Hence, we expect the following ranking of fairness ratings:

H2a: Fairness ratings are highest in the HUMAN treatment and lowest in the MACHINE treatment.

H2b: Fairness ratings for the HIGH treatment and the LOW treatment lie in between, with fairness ratings being higher in the HIGH treatment than in the LOW treatment.

However, even if participants generally prefer human decision-making, they might also prefer a procedure that processes as much information as possible without sacrificing human control. In our study, this is the *HIGH* treatment. We therefore pose the contradicting hypothesis:

H2c: Fairness ratings in the HIGH treatment will exceed fairness ratings in all other treatments.

We test our treatments in different contexts of public decision-making. While we do not have clear predictions on how context and procedure might interact in our study, we know from the literature that the perceived relative fairness of machine-based and human decision procedures might change with context.

Finally, in exploratory analyses, we study how fairness ratings in the different treatments are affected by the perceived accuracy of the procedures, and whether socio-demographic characteristics such as race, gender, and political orientation are predictive of fairness ratings.

4. Results

Our main research question pertains to the effects of different forms of algorithmic assistance in public-sector decision-making. These results are captured by the between-subjects treatment differences in our experiment. In Subsection 4.1, we report analyses of these treatment differences on the pooled data over all scenarios. These analyses also include discussions of (i) the relationship between the perceived accuracy of the different procedures and procedural fairness, (ii) the role of socio-demographic characteristics, and (iii) the impact of political affiliation on fairness ratings. In Subsection 4.2, we delve deeper into the context-specific effects of the four treatments in the three different scenarios. In all our analyses, we focus on fairness ratings of the different decision-making procedures.[10]

4.1. Overall Treatment Effects

According to our experimental design, each participant responds to the same treatment (in a different scenario) at three points in time. We observe that the participants' first response differs from the other two responses (average fairness ratings over all treatments at position $1/2/3 = 5.05/4.85/4.78$). However, these differences in fairness ratings seem to be mere level effects resulting from the timing of the response. There appears to be no systematic difference between

[10] Procedural preferences do not seem to differ between cases with personal involvement and cases applied to the general population. Hence, we relegate the summary analysis of our results on procedural preferences regarding the involvement of oneself or others to Appendix A.

responses at different points in time related to the treatments.[11] Treatment-specific order effects being absent, we therefore run our analyses at the group level on the data pooled from all responses across time.

Treatment Effects

As can be seen in Figure 3, fairness ratings are highest in the *HIGH* treatment ($M = 5.202$) with a human-computer interaction and high human control over the decision-making procedure. By contrast, participants judge the *MACHINE* treatment ($M = 4.638$) as the least fair. Participants relatively dislike when human decision-makers totally relinquish decision control. The *HUMAN* ($M = 4.887$) and the LOW treatment ($M = 4.842$) with human-computer interaction and low human control are in between. In sum, however, fairness ratings are relatively high in all treatments. More specifically, it is worth noting that fairness ratings are above the midpoint of the scale in all treatments, which suggests that all decision-making procedures seem to be acceptable in terms of procedural fairness.

[11] We refer to the Appendix for analyses of potential differential effects of the point in time of the response according to treatment (Table 4 in Appendix B). We only find a marginally significant difference between the effects of the position of the response in the *HIGH* treatment compared to the *MACHINE* treatment.

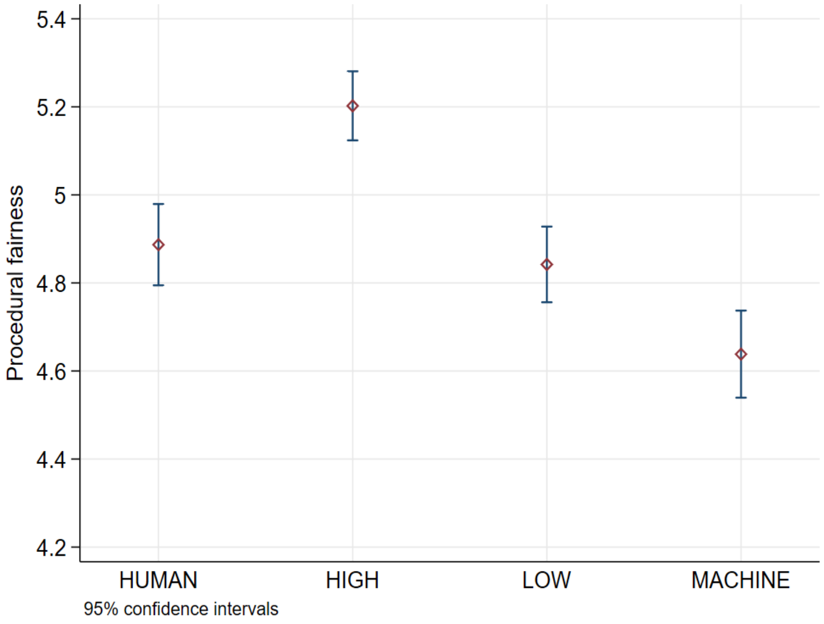
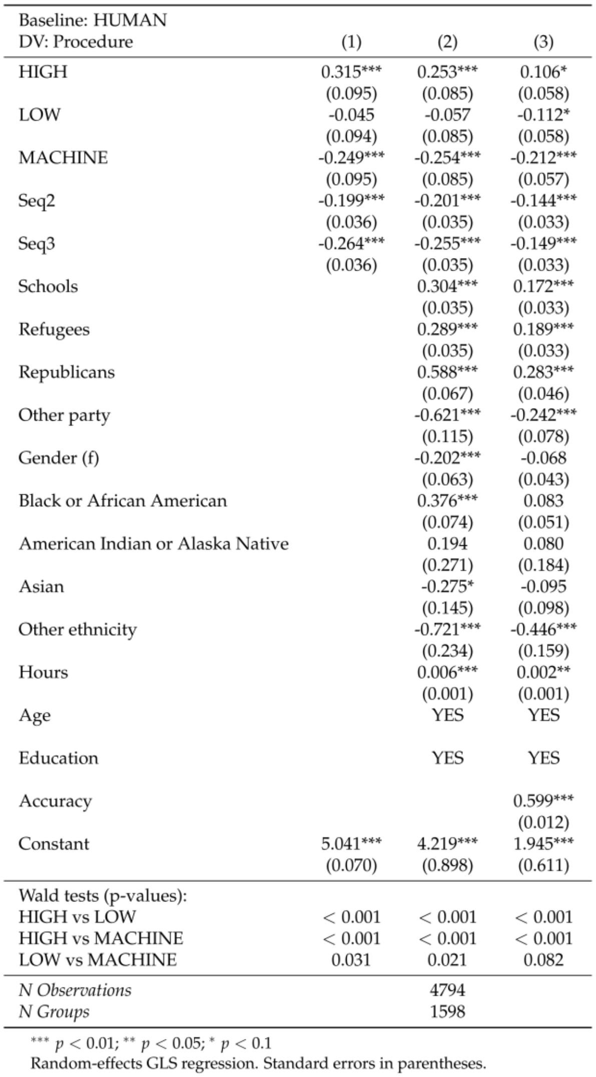


Figure 3: Procedural fairness across all scenarios

Overall, treatment differences are statistically significant according to non-parametric *Mann-Whitney U* (MWU) tests. Fairness ratings in the *HIGH* treatment are significantly higher than fairness ratings in all other treatments ($p < 0.001$, MWU). Participants seem to value the importance of human involvement in the decision-making process. Consequently, the purely algorithmic decision procedure in the *MACHINE* treatment yields significantly lower fairness ratings than all other treatments ($p < 0.5$, MWU). The difference in fairness ratings between the *HUMAN* and the *LOW* treatment, however, does not reach statistical significance ($p = 0.168$, MWU). This might support the interpretation that people accept a certain delegation of decisions to an algorithmic decision aid. Even a procedure in which the human decision-maker regularly just follows the machine advice yields similar fairness ratings as a purely human decision procedure.

Random-effects generalized least squares regression models confirm these results. All model specifications are displayed in Table 1. In Model 1, we regress fairness ratings on treatment dummies and dummies for the decision point in time. We control for the different scenarios and for participants' socio-demographic characteristics elicited in the post-experimental survey in Model 2. To be specific, we include dummy variables for the scenarios, as well as participants' gender, ethnicity, age, education, and their political preferences. We also include the self-reported amount of time the participants in our sample spend on paid online tasks. To explore possible explanations for our treatment effects, we add the participants' accuracy ratings to the regression estimation in Model 3.

Table 1: Treatment effects on procedural fairness across scenarios



Baseline: HUMAN DV: Procedure	(1)	(2)	(3)
HIGH	0.315***	0.253***	0.106*
	(0.095)	(0.085)	(0.058)
LOW	-0.045	-0.057	-0.112*
	(0.094)	(0.085)	(0.058)
MACHINE	-0.249***	-0.254***	-0.212***
	(0.095)	(0.085)	(0.057)
Seq2	-0.199***	-0.201***	-0.144***
	(0.036)	(0.035)	(0.033)
Seq3	-0.264***	-0.255***	-0.149***
	(0.036)	(0.035)	(0.033)
Schools		0.304***	0.172***
		(0.035)	(0.033)
Refugees		0.289***	0.189***
		(0.035)	(0.033)
Republicans		0.588***	0.283***
		(0.067)	(0.046)
Other party		-0.621***	-0.242***
		(0.115)	(0.078)
Gender (f)		-0.202***	-0.068
		(0.063)	(0.043)
Black or African American		0.376***	0.083
		(0.074)	(0.051)
American Indian or Alaska Native		0.194	0.080
		(0.271)	(0.184)
Asian		-0.275*	-0.095
		(0.145)	(0.098)
Other ethnicity		-0.721***	-0.446***
		(0.234)	(0.159)
Hours		0.006***	0.002**
		(0.001)	(0.001)
Age		YES	YES
Education		YES	YES
Accuracy			0.599***
			(0.012)
Constant	5.041***	4.219***	1.945***
	(0.070)	(0.898)	(0.611)
Wald tests (p-values):			
HIGH vs LOW	< 0.001	< 0.001	< 0.001
HIGH vs MACHINE	< 0.001	< 0.001	< 0.001
LOW vs MACHINE	0.031	0.021	0.082
N Observations		4794	
N Groups		1598	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$
Random-effects GLS regression. Standard errors in parentheses.

With the *HUMAN* treatment as the reference category, we observe that the reported differences of the fairness ratings between our treatments are robust to the inclusion of all control variables added in Model 2. The coefficients for the *HIGH* treatment dummy and the *MACHINE* treatment dummy are positive and negative, respectively, and turn out to be statistically significant. The coefficient for the *LOW* treatment, in contrast, is close to zero and statistically insignificant. Wald tests, run after the estimation of Model 2, confirm the treatment differences between the *HIGH* treatment and either the *LOW* or the *MACHINE* treatment, as well as between the latter treatments ($p < 0.05$). This leads to the following main results of our study:

Result 1: *Fairness ratings are responsive to different degrees of human involvement in the decision procedures (in support of H1).*

Result 2: *A human-machine interaction with high human involvement is judged as fairer than the decision procedures in all other treatments (in support of H2c).*

Result 3: *Purely machine-based decision procedures receive the lowest fairness scores of all procedures (in partial support of H2a).*

Result 4: *Purely human decision-making and human-machine interactions with low human involvement are perceived as equally fair (not hypothesized).*

On further inspection of the control variables included in Model 2, we find that people who identify as Republicans show higher fairness ratings than people who identify as Democrats. Participants who identify as neither Republican nor Democrat report significantly lower fairness evaluations than Democrats. Moreover, the coefficient of the Gender dummy also turns out significant, with women reporting lower fairness evaluations than men. We also observe a positive correlation between identifying as African American and fairness ratings. We discuss these findings on the fairness ratings of several subgroups in more details below.

Decision Accuracy

In Model 3, we observe a significant effect of the participants' accuracy assessments on fairness ratings. Controlling for accuracy considerably changes the coefficients of our treatment dummies. However, coefficients for the *HIGH* and *MACHINE* treatment dummies keep their sign and remain (marginally) significant (*HIGH*: $p < 0.064$, *MACHINE*: $p < 0.001$), whereas the coefficient for the *LOW* treatment is now clearly negative and marginally significant ($p = 0.053$). Post-regression Wald tests confirm the further treatment differences, also after controlling for expected accuracy.

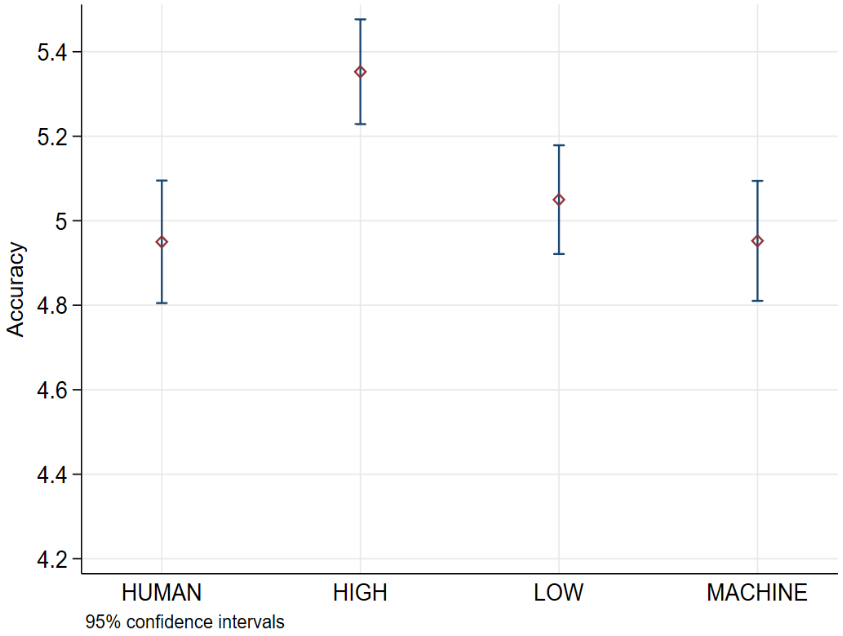


Figure 4: Perceived accuracy across all scenarios

These results lead to the interpretation that people largely seem to prefer the human-computer interaction in the *HIGH* treatment because they think this specific interaction leads to a more accurate result. However, while a large part of the fairness gain from a human-machine interaction with high human involvement seems to stem from higher expected accuracy of these procedures, the relative dislike of a purely algorithmic decision in the *MACHINE* treatment is practically not affected by the inclusion of the accuracy assessments (coefficients for the *MACHINE* treatment are of roughly equal size in Model 2 and Model 3 and remain significant in both models). The difference between the HIGH and the LOW treatment remains significant after controlling for accuracy ($p = 0.001$, Wald test).

This suggests that participants are sensitive to variations in the degree of human involvement and form their accuracy assessments and fairness ratings based on the degree of human involvement in algorithmically assisted decision procedures. Moreover, we see these results as an indication that the rejection of purely algorithmic decisions is not merely driven by the expectation that algorithms make more mistakes. More generally, it seems that a combination of algorithmic and human decision inputs is perceived to produce more accurate factual assessments. As can be seen in Figure 4, high human involvement in the algorithmic decision procedure, as in the *HIGH* treatment, is likely to foster the perceived accuracy of the procedure, as compared to all other conditions ($p = 0.001$, MWU). Participants report no difference in expected accuracy between the *HUMAN* and the *MACHINE* treatment ($p = 0.223$, MWU). This seems noteworthy because arguably participants overestimate the prediction capability of human decision-makers or, to put it the other way around, underestimate the capability of algorithmic prediction, as a long-standing literature indicates that, overall, statistical models outperform humans in prediction tasks (Meehl, 1954; Grove et al., 2000; Kleinberg et al., 2017).

Previous findings suggest that part of the effect of fair procedures on perceived legitimacy is mediated by the belief that fair procedures yield more accurate outcomes (Tyler & Sevier, 2014). Our results so far corroborate that accuracy may play a similar role in people's fairness assessments of different decision procedures in human-machine interactions. In the literature, Wang (2018) has shown that accuracy affects fairness judgments. Studying the fairness of specific features used in algorithmic predictions, Grgić-Hlača et al. (2018b) find that fairness ratings increase when it is assumed that a feature enhances the accuracy of the prediction. In the study of Albach and Wright (2021), how a specific feature contributes to the accuracy of the decision emerges as people's main concern when they form their fairness assessment of the use of this feature in an algorithmic decision-making process.[12]

To explore the conjecture further that the perceived accuracy of the procedure mediates fairness ratings in the context of our study, we conduct a mediation analysis to measure the *direct effect* of our treatments (x_i) on fairness ratings (y_i) and the *indirect effect* of our treatments on fairness ratings through accuracy assessments as a mediator (z_i). These effects can be estimated in a structural equation model, with the treatment effect on the mediator given by:

$$z_{it} = \alpha_0 + \alpha_x x_{it} + u_{it} + \varepsilon_i \ ,$$

with u_{it} denoting the residual error between individuals and ε_i denoting the individual-specific error. The full structural equation model can be specified as follows:

$$y_{it} = \beta_0 + \beta_x x_{it} + \beta_z z_{it} + +u_{it} + \varepsilon_i \ .$$

The direct treatment effect is given by β_x, denoting the pathway from treatment to fairness ratings while controlling for accuracy assessments. The indirect treatment effect is given by $\gamma_I = \alpha_x \cdot \beta_z$, denoting the pathway through accuracy assessments.[13]

[12] For a non-legal setting, Yin et al. (2019) report experimental evidence that the stated accuracy of a machine-learning model may affect self-reported trust in the model.

[13] The total treatment effect is given by $\gamma_T = \beta_x + \alpha_x \cdot \beta_z$ and already reported, for slightly different model specifications, in Table 1.

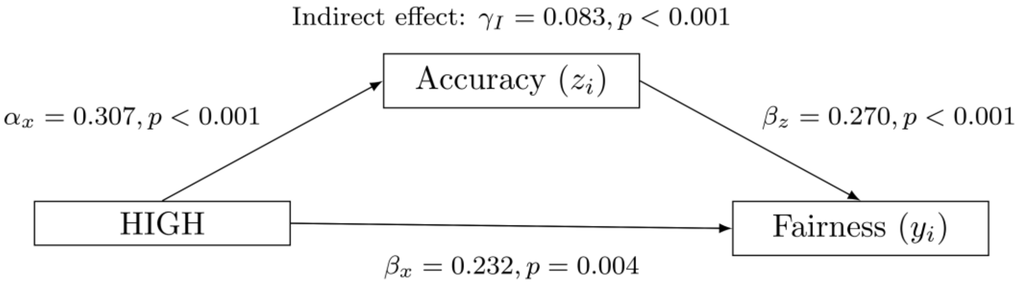

Indirect effect: $\gamma_I = 0.083, p < 0.001$
Accuracy (z_i)
$\alpha_x = 0.307, p < 0.001$
$\beta_z = 0.270, p < 0.001$
HIGH
Fairness (y_i)
$\beta_x = 0.232, p = 0.004$

Figure 5: Mediation analysis HIGH vs. HUMAN

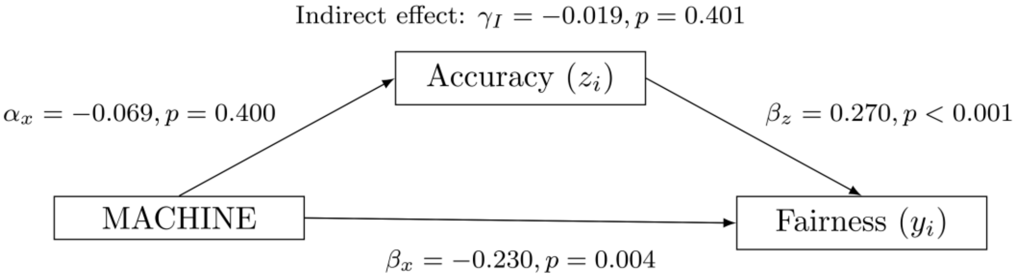

Indirect effect: $\gamma_I = -0.019, p = 0.401$
Accuracy (z_i)
$\alpha_x = -0.069, p = 0.400$
$\beta_z = 0.270, p < 0.001$
MACHINE
Fairness (y_i)
$\beta_x = -0.230, p = 0.004$

Figure 6: Mediation analysis MACHINE vs. HUMAN

The results of our structural equation models show that a considerable part of the *HIGH* treatment effect compared to the *HUMAN* treatment follows the indirect path through accuracy assessments (Figure 5). In the *MACHINE* treatment, by contrast, we observe no significant indirect effect mediated by accuracy assessments (Figure 6). This corroborates our conjecture that the decrease of fairness ratings observed for purely algorithm-based decision procedures is mostly driven by cognitive or motivational effects that are unrelated to perceived accuracy.

***Result 5**:* *In parts, the HIGH treatment is judged as fairer than the HUMAN treatment because it is perceived as more accurate. The relative dislike of the MACHINE treatment is not affected by accuracy assessments (not hypothesized).*

Ethnicity and Gender

A further observation from Model 3 in Table 1 is that the coefficients for the dummy variables for African American ethnicity and for gender are much smaller and no longer significant once we control for accuracy assessments ($p = 0.102$ and $p = 0.111$, respectively). This indicates that the higher fairness ratings of people identifying as African American and of men compared to women are also in parts driven by the perceived accuracy of the procedure. Indeed, accuracy assessments of participants identifying as African American ($M = 5.652$) are significantly higher than average accuracy assessments of participants belonging to all other ethnic groups ($M = 4.755$,$p < 0.001$, MWU). We also find that female participants ($M = 4.796$) express significantly lower accuracy ratings than male participants ($M = 5.094$, $p < 0.001$, MWU).

Political Affiliation

While the effects of ethnicity and gender vanish once we control for accuracy assessments, the effect of political affiliation seems more robust to the inclusion of all our covariates (Model 3). With Democrats as the reference category, the coefficient for Republicans remains consistently positive throughout all model specifications, whereas we observe a consistently negative effect of identifying with a political ideology beyond the bipartisan Democrat-Republican spectrum. Republicans thus tend to view all procedures as fairer than Democrats, while participants who identify with other parties tend to express lower fairness ratings than the two parties dominating the political landscape in the US (Figure 7). These differences are significant both for the comparison between Democrats and Republicans ($p < 0.001$) and for the comparison between Democrats and participants who identify with other parties ($p = 0.002$).

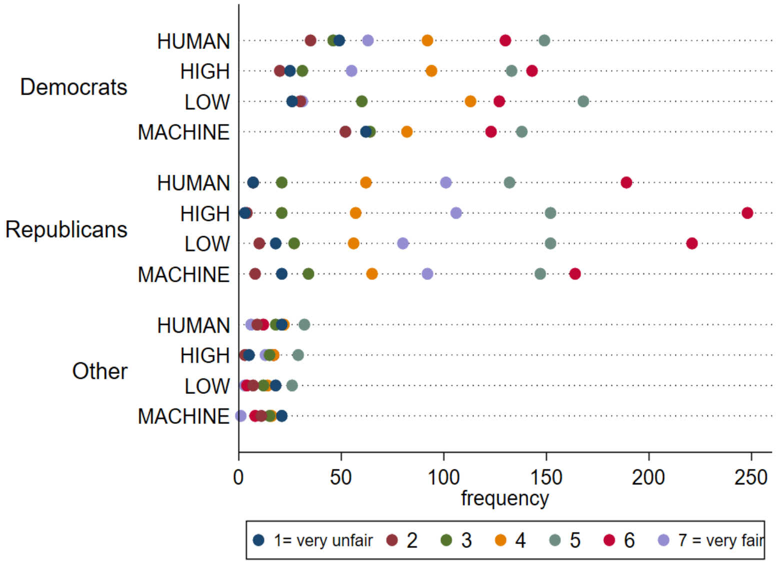


Figure 7: Effect of political affiliation

4.2. Scenario-Specific Effects

Descriptively, the overall pattern of the aggregated results is also present if we look at the treatments in the three scenarios individually.[14] Fairness ratings in the treatments for each scenario can be seen in Figure 8. In all three scenarios, fairness ratings are highest in the *HIGH* treatment and lowest in the *MACHINE* treatment. In the school-admissions and the refugee-matching scenarios, the fairness ratings of the other two treatments are in between, with the *HUMAN* treatment being assessed as (slightly) fairer than the *LOW* treatment.

[14] As mentioned before, each participant answered the fairness question in the same treatment in three different scenarios. The effects of the timing of the decision seem to be generally unaffected by the different scenarios. In Table 4 in Appendix B, we report a random-effects generalized least squares regression model, in which all interactions of the decision point in time and the scenarios turn out non-significant, with the exception of the refugee-matching scenario, in which the fact that the scenario was presented last produces a (marginally) significant more negative effect than the two other treatments.

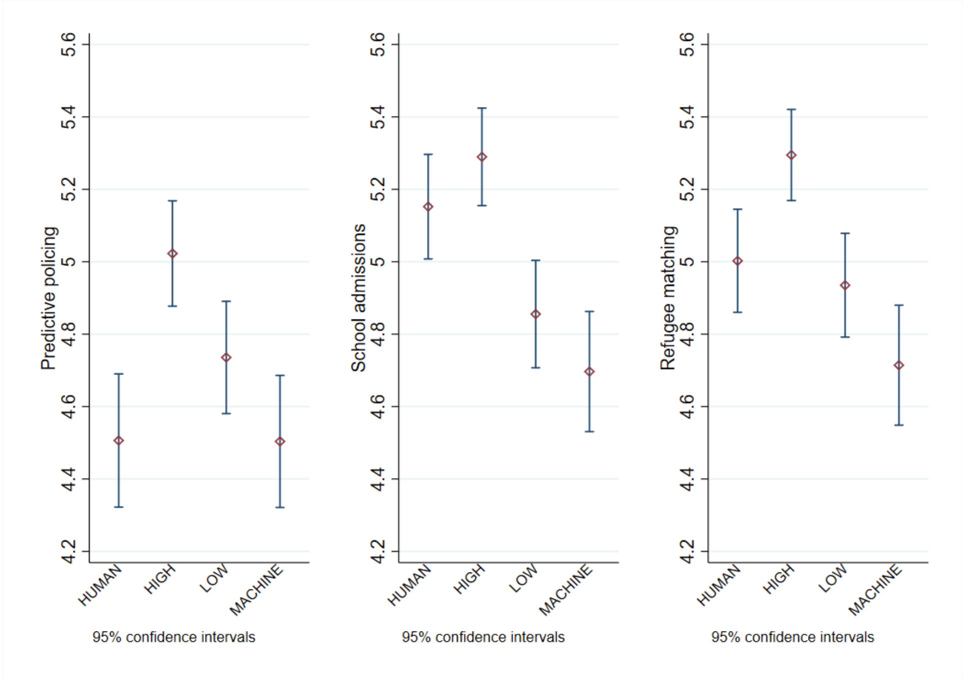


Figure 8: Procedural fairness by scenario

The predictive-policing scenario stands out in this regard, as decisions by a human police officer are considered less fair than decisions by human decision-makers in the other two scenarios.[15] Our analysis suggests that the fairness-enhancing effect of a human decision-maker is entirely captured by the school-admissions and refugee-matching context. Overall, there seems to be a context-specific difference between human police officers and other public officials.

Accordingly, in the predictive-policing scenario, we find that the average fairness ratings in the $HUMAN$ ($M = 4.506$) and $MACHINE$ treatment ($M = 4.504$) are virtually identical ($p = 0.883$, MWU). Moreover, we do not find a significant difference either between the LOW ($M = 4.736$)

[15] This can be shown in a random-effects generalized least squares regression model estimating treatment effects on fairness ratings, with the *MACHINE* treatment and the predictive-policing scenario as reference categories (Table 5 in Appendix C). We observe a significant effect of all treatments both in our base specification (Model 1) and in our specification including dummies as for the school admissions and the refugee-matching scenarios as controls (Model 2). When including an interaction term for treatment and scenario, however, the coefficient for the *HUMAN* treatment is no longer significant, whereas we observe a significant interaction between the *HUMAN* treatment and the school-admissions and the refugee-matching scenarios (Model 3).

and the *MACHINE* treatment ($p = 0.199$, MWU) or between the *LOW* and the *HUMAN* treatment ($p = 0.288$, MWU). Yet we observe significantly higher fairness ratings in the *HIGH* treatment ($M = 5.023$) than in all other treatments ($p < 0.05$, MWU). We interpret these results as evidence of relatively strong fairness preferences for hybrid predictive-policing procedures involving the combined input of humans and algorithms.

The school-admissions and the refugee-matching scenario look much more similar, with the *HIGH* treatment being consistently perceived as the fairest and the *HUMAN* treatment performing consistently better in terms of fairness than the *MACHINE* treatment across both scenarios.

In the school-admissions scenario, fairness ratings are highest in the *HUMAN* ($M = 5.152$) and the *HIGH* ($M = 5.290$) treatment, with both treatments being rather close to each other ($p = 0.211$, MWU). The *HIGH* treatment yields significantly higher fairness ratings than the *LOW* ($M = 4.855$) and *MACHINE* treatment ($M = 4.697, p < 0.001$ respectively, MWU). Also, the *HUMAN* treatment leads to significantly higher fairness ratings than the *LOW* ($p = 0.004$, MWU) and the *MACHINE* treatment ($p < 0.001$, MWU). The pronounced difference between our treatments with strong human involvement and the other two (more algorithmic) treatments points at the particular importance of human oversight in areas as sensitive as school admissions. The markedly positive effect of our *HUMAN* treatment may also be due to the fact that, unlike in the other scenarios, a group – the school-admissions board – rather than individuals made the decision.

In the refugee-matching scenario, by contrast, the *HIGH* treatment ($M = 5.295$) produces significantly higher fairness ratings than all other treatments ($p < 0.004$, MWU). However, fairness ratings differ neither between the *HUMAN* ($M = 5.002$) and the *LOW* treatment ($M = 4.935$, $p = 0.582$, MWU) nor between the *LOW* and the *MACHINE* treatment ($M = 4.714$, $p = 0.165$, MWU). Moreover, even when comparing the *HUMAN* and the *MACHINE* treatment, we only find a marginally positive effect of an entirely human refugee-matching procedure ($p = 0.055$, MWU). While a procedure based on human-computer interaction and high human control is viewed as bolstering procedural fairness, the degree of human involvement does not seem to matter much when it comes to refugee-matching. This may be because issues of distributive justice or participatory rights of those affected by the decision are less salient in refugee-matching procedures than in other contexts.

5. Conclusion

In this article, we report experimental evidence on the importance of human involvement in algorithmically assisted public-sector decision-making for fairness perceptions. We find for several application contexts that procedures are perceived as fairest when an algorithmic decision aid is accompanied by high human involvement in the decision-making procedure. Arguably, this is the case to a large extent because people expect these procedures to be the most accurate. By contrast, we observe that purely algorithmic decisions are consistently judged as least fair. This dislike seems to be largely independent of the perceived accuracy of the decision-making procedure. Therefore, while perceived accuracy matters for fairness perceptions in our experiment, it cannot fully explain people's dislike for purely algorithmic decision-making. This is in line with previous findings on accuracy and procedural fairness in bail decisions (Wang, 2018).

While a high level of human involvement boosts the procedural fairness of algorithmic assistance, it counteracts the efficiency promises of algorithmic decision aids. In our treatment with high human involvement, human and algorithmic decision-making always coincide. There is no real substitution of human decision-making by the algorithm. However, our findings lend support to the view that decision-making procedures with reduced human involvement might yield similar fairness perceptions as purely human decision-making procedures. This suggests that moving from the status quo of public decision-making by humans towards mainly algorithmic decision-making procedures may be less disruptive in terms of procedural fairness than the law and policy debate sometimes suggests. In our treatment with low human involvement, the decision is usually based on the algorithmic advice alone, with the human decision-maker only sometimes engaging in a personal assessment of the facts. This leads to largely similar fairness ratings than an entirely human decision-making procedure. Hence, while human involvement matters to people, they are relatively open to moderate degrees of decision delegation to a machine.

Our findings on treatment differences come with caveats, of course. One limitation of our study stems from the fact that, in all our treatments with human involvement, the human decision-maker at least theoretically retains final control. The human decision-maker can reverse every decision by the algorithm. The delegation of decision power to the machine in our treatment with low human involvement is therefore only factual. Human decision-makers de facto forgo the opportunity to evaluate the facts of the case, but they are not legally obliged to refrain from performing their own assessment.

Moreover, treatment differences are in some instances sensitive to the decision context. We find noteworthy differences between the three scenarios for predictive policing, school admissions, and refugee-matching. For example, assessments of human decisions considerably vary across contexts, with the predictive-policing scenario showing considerably lower fairness ratings for a human decision-maker as compared to the other two treatments. This difference may reflect a general loss of trust in human police officers in light of repeated abuses of police authority and increasing public awareness of police brutality, such as the murder of George Floyd in 2020. Differences between scenarios are important because they indicate that there may be no *one-size-*

fits-all solution for the use of algorithms in public-sector decision-making. For example, fairness perceptions of human decisions are rather high in the school-admissions context. While this may be due to the perceived importance of school admissions or the fact that the admissions decision is made by a collective in this scenario, our experiment is not designed to generate data in support of these interpretations. It is up to future research to explore the optimal mix of human and algorithmic involvement in decision-making procedures for specific policy fields.

References

Acharya, Avidit, Kirk Bansak, & Jens Hainmueller. 2022. Combining Outcome-Based and Preference-Based Matching: A Constrained Priority Mechanism. *Political Analysis* 30 (1), 89–12.

Ahani, Narges, Tommy Andersson, Alessandro Martinello, Alexander Teytelboym, & Andrew C. Trapp. 2021. Placement Optimization in Refugee Resettlement. *Operations Research* 69 (5), 1468–1486.

Albach, Michele, & James R. Wright. 2021. The Role of Accuracy in Algorithmic Process Fairness Across Multiple Domains. In *Proceedings of the 22nd ACM Conference on Economics and Computation (EC'21)*, 29–49.

Araujo, Theo, Natali Helberger, Sanne Kruikemeier, & Claes H. De Vreese. 2020. In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence. *AI & SOCIETY* 35 (3), 611–23.

Bansak, Kirk, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, & Jeremy Weinstein. 2018. Improving Refugee Integration Through Data-Driven Algorithmic Assignment. *Science* 359 (6373), 325–29.

Barocas, Solon, Moritz Hardt, & Arvind Narayanan. 2021. *Fairness in Machine Learning: Limitations and Opportunities*.

Becker, Gary S. 1968. Crime and Punishment: An Economic Approach. *Journal of Political Economy* 76, 169–217.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, & Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50 (1), 3–44.

Binns, Reuben. 2022. Human Judgment in Algorithmic Loops: Individual Justice and Automated Decision-Making. *Regulation & Governance* 16 (1), 197–211.

Blount, Sally. 1995. When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences. *Organizational Behavior and Human Decision Processes* 63 (2), 131–44.

Casey, Anthony J., & Anthony Niblett. 2017. The Death of Rules and Standards. *Indiana Law Journal* 92 (4), 1401–47.

Chen, Benjamin Minhao, Alexander Stremitzer, & Kevin Tobia. 2022. Having Your Day in Robot Court. *Harvard Journal of Law & Technology* 36, forthcoming.

Chouldechova, Alexandra. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5 (2), 153–63.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, & Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*, 797–806.

Cummings, Mary L. 2017. Automation Bias in Intelligent Time Critical Decision Support Systems. In Don Harris, & Wen-Chin Li (Eds.), *Decision Making in Aviation*, 289–294. Routledge.

Dietvorst, Berkeley J., Joseph P. Simmons, & Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General* 144, 114–26.

Dodge, Jonathan, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, & Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*, 275–85.

Falk, Armin, Ernst Fehr, & Urs Fischbacher. 2008. Testing Theories of Fairness – Intentions Matter. *Games and Economic Behavior* 62 (1), 287–303.

Green, Ben, & Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, 90–99.

Grgić-Hlača, Nina, Elissa M. Redmiles, Krishna P. Gummadi, & Adrian Weller. 2018a. Human Perceptions of Fairness in Algorithmic Decision Making. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web (WWW '18)*, 903–12.

Grgić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, & Adrian Weller. 2018b. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1), 51–60.

Grove, William M., David H. Zald, Boyd S. Lebow, Beth E. Snitz, & Chad Nelson. 2000. Clinical Versus Mechanical Prediction: A Meta-Analysis. *Psychological Assessment* 12 (1), 19–30.

Güth, Werner, Rolf Schmittberger, & Bernd Schwarze. 1982. An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization* 3 (4), 367–88.

Harrison, Galen, Julia Hanson, Christine Jacinto, Julio Ramirez, & Blase Ur. 2020. An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, 392–402.

Hellman, Deborah. 2020. Measuring Algorithmic Fairness. *Virginia Law Review* 106 (4), 811–66.

Horton, John J., David G. Rand, & Richard J. Zeckhauser. 2011. The Online Laboratory: Conducting Experiments in a Real Labor Market. *Experimental Economics* 14, 399–425.

Huq, Aziz Z. 2020a. A Right to a Human Decision. *Virginia Law Review* 106 (3), 611–88.

———. 2020b. “Constitutional Rights in the Machine-Learning State.” *Cornell Law Review* 105 (7), 1875–1954.

Imai, Kosuke, Zhichao Jiang, James Greiner, Ryan Halen, & Sooahn Shin. 2021. Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment. *arXiv:2012.02845v2 [Cs.CY]*.

Inaba, Misato, Yumi Inoue, Satoshi Akutsu, Nobuyuki Takahashi, & Toshio Yamagishi. 2018. Preference and Strategy in Proposer’s Prosocial Giving in the Ultimatum Game. *PLoS ONE* 13 (3), e0193877.

Joh, Elizabeth E. 2016. The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing. *Harvard Law & Policy Review* 10 (1), 15–42.

Kearns, Michael, & Aaron Roth. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, & Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions. *Quarterly Journal of Economics* 133 (1), 237–93.

Kleinberg, Jon, Sendhil Mullainathan, & Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 67 (43), 1–43.

Lee, Min Kyung, & Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '17)*, 1035–48.

Lee, Min Kyung, Anuraag Jain, Hae Jin Cha, Shashank Ojha, & Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. In *Proceedings of the ACM on Human-Computer Interaction*, 3 (CSCW), 1–26.

Lind, E. Allan, & Tom R. Tyler. 1988. *The Social Psychology of Procedural Justice*. Plenum Press.

Logg, Jennifer M., Julia A. Minson, & Don A. Moore. 2019. Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes* 151, 90–103.

Manzey, Dietrich, Juliane Reichenbach, & Linda Onnasch. 2012. Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making* 6(1), 57–87.

Marcinkowski, Frank, Kimon Kieslich, Christopher Starke, & Marco Lünich. 2020. Implications of AI (Un-)Fairness in Higher Education Admissions: The Effects of Perceived AI (Un-)Fairness on Exit, Voice and Organizational Reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, 122–30.

Meehl, Paul E. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* University of Minnesota Press.

Miller, Susan M., & Lael R. Keiser. 2021. Representative Bureaucracy and Attitudes Toward Automated Decision Making. *Journal of Public Administration Research and Theory* 31 (1), 150–65.

Muratov, Eugene, Margaret Lewis, Denis Fourches, Alexander Tropsha, & Wendy C. Cox. 2017. Computer-Assisted Decision Support for Student Admissions Based on Their Predicted Academic Performance. *American Journal of Pharmaceutical Education* 81 (3), 1–9.

Nagtegaal, Rosanna. 2021. The Impact of Using Algorithms for Managerial Decisions on Public Employees' Procedural Justice. *Government Information Quarterly* 38 (1), 101536.

Offerman, Theo. 2002. Hurting Hurts More Than Helping Helps. *European Economic Review* 46 (8), 1423–37.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, & Sendhil Mullainathan. 2020. An Economic Perspective on Algorithmic Fairness. *AEA Papers and Proceedings* 110, 91–95.

Roth, Alvin E. 1995. Bargaining Experiments. In John Kagel & Alvin E. Roth (Eds.), *Handbook of Experimental Economics*, 253–348. Princeton University Press.

Simmons, Ric. 2018. Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System. *University of California Davis Law Review* 52 (2), 1067–1118.

Starke, Christopher, Janine Baleis, Birte Keller, & Frank Marcinkowski. 2021. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. *arXiv:2103.12016v1 [Cs.HC]*.

Tyler, Tom R. 2003. Procedural Justice, Legitimacy, and the Effective Rule of Law. *Crime and Justice* 30, 283–357.

———. 2006. *Why People Obey the Law*. Princeton University Press.

Tyler, Tom R., & Justin Sevier. 2014. How Do the Courts Create Popular Legitimacy: The Role of Establishing the Truth, Punishing Justly, and/or Acting through Just Procedures. *Albany Law Review* 77(3), 1095–1137.

Walker, Laurens, E. Allan Lind, & John Thibaut. 1979. The Relation Between Procedural and Distributive Justice. *Virginia Law Review* 65 (8), 1401–1420.

Wang, A. J. 2018. Procedural Justice and Risk-Assessment Algorithms. *Yale Law School, Working Paper*.

Yaghini, Mohammad, Hoda Heidari, & Andreas Krause. 2021. A Human-in-the-Loop Framework to Construct Context-Dependent Mathematical Formulations of Fairness. *arXiv:1911.03020v2 [Cs.AI]*.

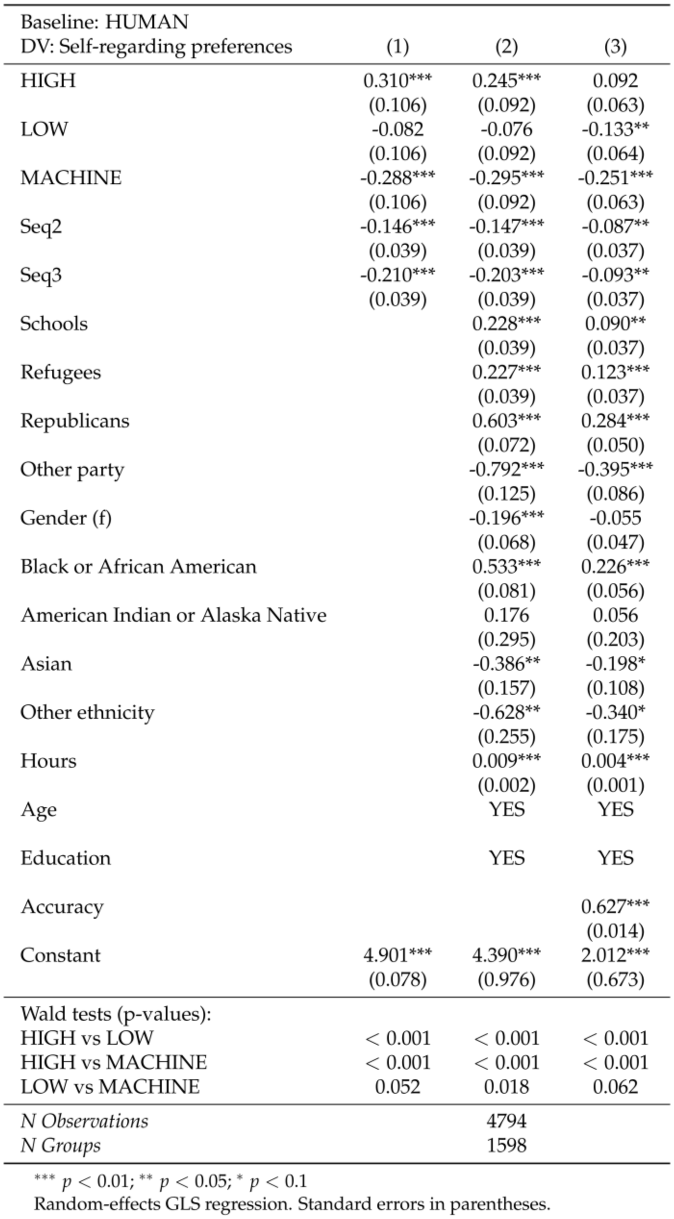
Yin, Ming, Jennifer Wortman Vaughan, & Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19)*, Paper No. 279, 1–12.

Appendix

A. Treatment Effects on Procedural Preferences

In this section, we present a summary analysis of our results on procedural preferences regarding the involvement of oneself (Table 2) or others (Table 3).

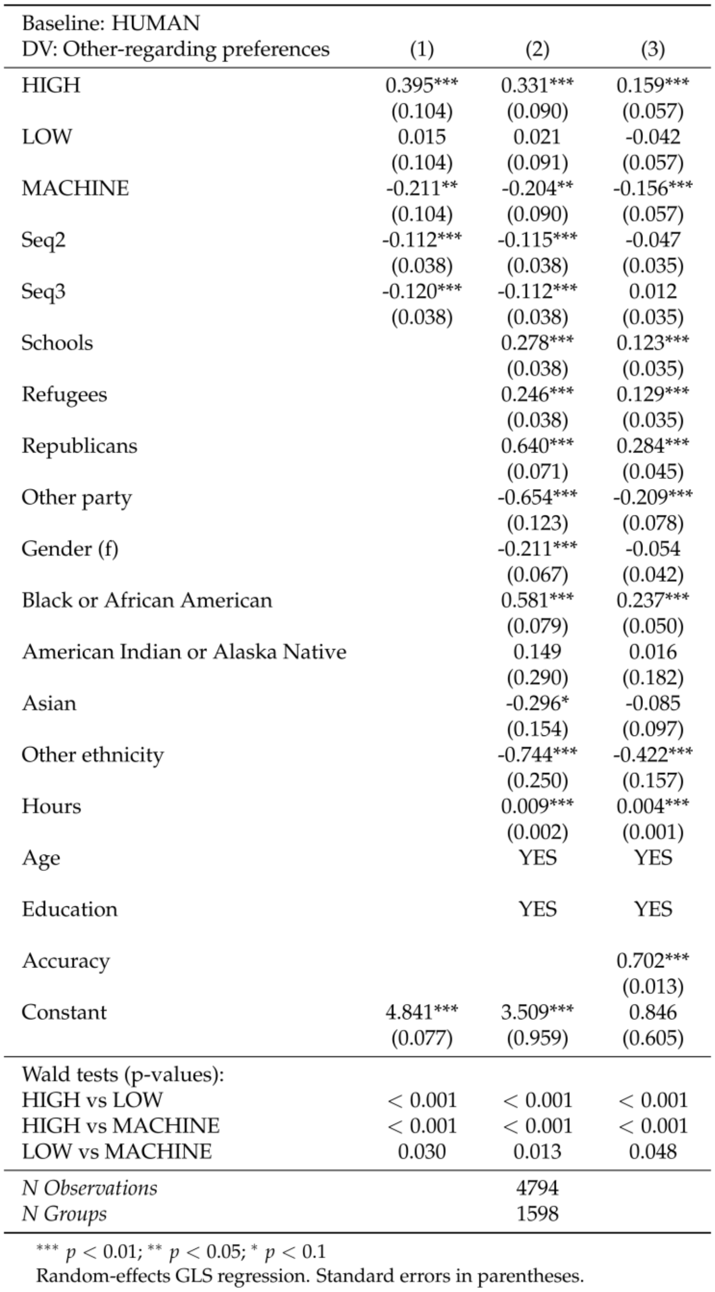
Table 2: Treatment effects on preferences regarding oneself across scenarios



Baseline: HUMAN DV: Self-regarding preferences	(1)	(2)	(3)
HIGH	0.310***	0.245***	0.092
	(0.106)	(0.092)	(0.063)
LOW	-0.082	-0.076	-0.133**
	(0.106)	(0.092)	(0.064)
MACHINE	-0.288***	-0.295***	-0.251***
	(0.106)	(0.092)	(0.063)
Seq2	-0.146***	-0.147***	-0.087**
	(0.039)	(0.039)	(0.037)
Seq3	-0.210***	-0.203***	-0.093**
	(0.039)	(0.039)	(0.037)
Schools		0.228***	0.090**
		(0.039)	(0.037)
Refugees		0.227***	0.123***
		(0.039)	(0.037)
Republicans		0.603***	0.284***
		(0.072)	(0.050)
Other party		-0.792***	-0.395***
		(0.125)	(0.086)
Gender (f)		-0.196***	-0.055
		(0.068)	(0.047)
Black or African American		0.533***	0.226***
		(0.081)	(0.056)
American Indian or Alaska Native		0.176	0.056
		(0.295)	(0.203)
Asian		-0.386**	-0.198*
		(0.157)	(0.108)
Other ethnicity		-0.628**	-0.340*
		(0.255)	(0.175)
Hours		0.009***	0.004***
		(0.002)	(0.001)
Age		YES	YES
Education		YES	YES
Accuracy			0.627***
			(0.014)
Constant	4.901***	4.390***	2.012***
	(0.078)	(0.976)	(0.673)
Wald tests (p-values):			
HIGH vs LOW	< 0.001	< 0.001	< 0.001
HIGH vs MACHINE	< 0.001	< 0.001	< 0.001
LOW vs MACHINE	0.052	0.018	0.062
N Observations		4794	
N Groups		1598	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$
Random-effects GLS regression. Standard errors in parentheses.

Table 3: Treatment effects on preferences regarding others across scenarios



Baseline: HUMAN DV: Other-regarding preferences	(1)	(2)	(3)
HIGH	0.395***	0.331***	0.159***
	(0.104)	(0.090)	(0.057)
LOW	0.015	0.021	-0.042
	(0.104)	(0.091)	(0.057)
MACHINE	-0.211**	-0.204**	-0.156***
	(0.104)	(0.090)	(0.057)
Seq2	-0.112***	-0.115***	-0.047
	(0.038)	(0.038)	(0.035)
Seq3	-0.120***	-0.112***	0.012
	(0.038)	(0.038)	(0.035)
Schools		0.278***	0.123***
		(0.038)	(0.035)
Refugees		0.246***	0.129***
		(0.038)	(0.035)
Republicans		0.640***	0.284***
		(0.071)	(0.045)
Other party		-0.654***	-0.209***
		(0.123)	(0.078)
Gender (f)		-0.211***	-0.054
		(0.067)	(0.042)
Black or African American		0.581***	0.237***
		(0.079)	(0.050)
American Indian or Alaska Native		0.149	0.016
		(0.290)	(0.182)
Asian		-0.296*	-0.085
		(0.154)	(0.097)
Other ethnicity		-0.744***	-0.422***
		(0.250)	(0.157)
Hours		0.009***	0.004***
		(0.002)	(0.001)
Age		YES	YES
Education		YES	YES
Accuracy			0.702***
			(0.013)
Constant	4.841***	3.509***	0.846
	(0.077)	(0.959)	(0.605)
Wald tests (p-values):			
HIGH vs LOW	< 0.001	< 0.001	< 0.001
HIGH vs MACHINE	< 0.001	< 0.001	< 0.001
LOW vs MACHINE	0.030	0.013	0.048
N Observations		4794	
N Groups		1598	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$
Random-effects GLS regression. Standard errors in parentheses.

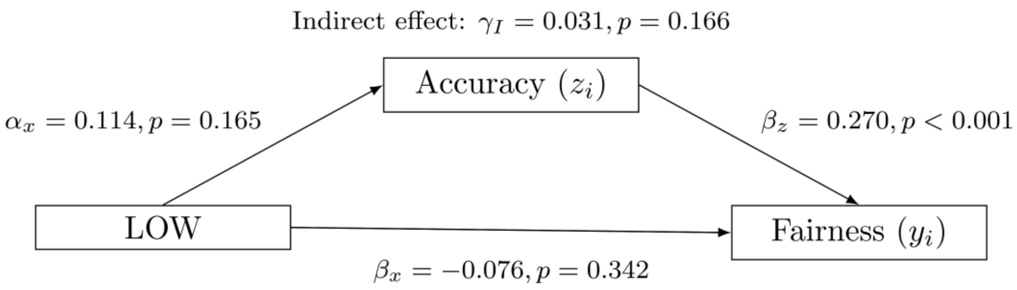
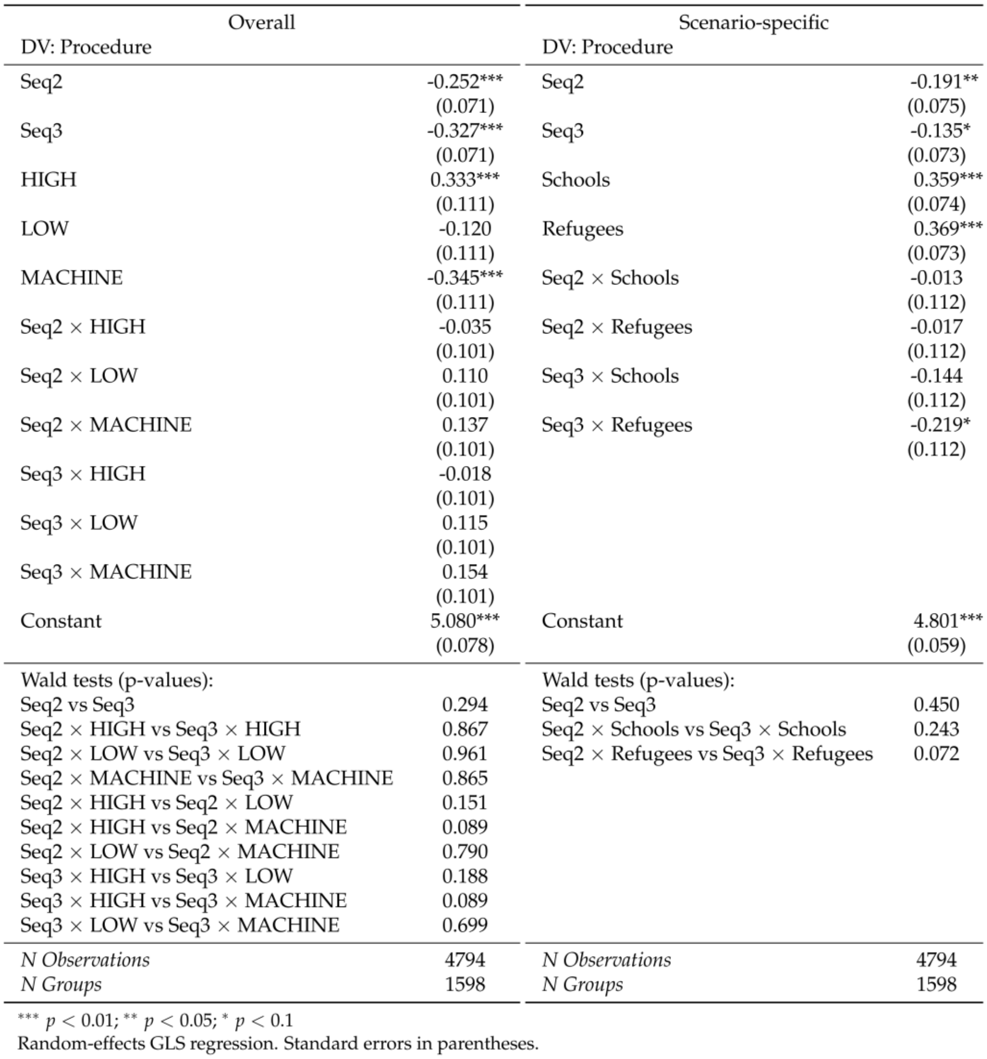

Indirect effect: $\gamma_I = 0.031, p = 0.166$
Accuracy (z_i)
$\alpha_x = 0.114, p = 0.165$
$\beta_z = 0.270, p < 0.001$
LOW
Fairness (y_i)
$\beta_x = -0.076, p = 0.342$

Figure 9: Mediation analysis LOW vs. HUMAN

B. Order Effects

In this section, we present an analysis of order effects pooled over all scenarios (Table 4, left column) and for each scenario (Table 4, right column).

Table 4: Order effects



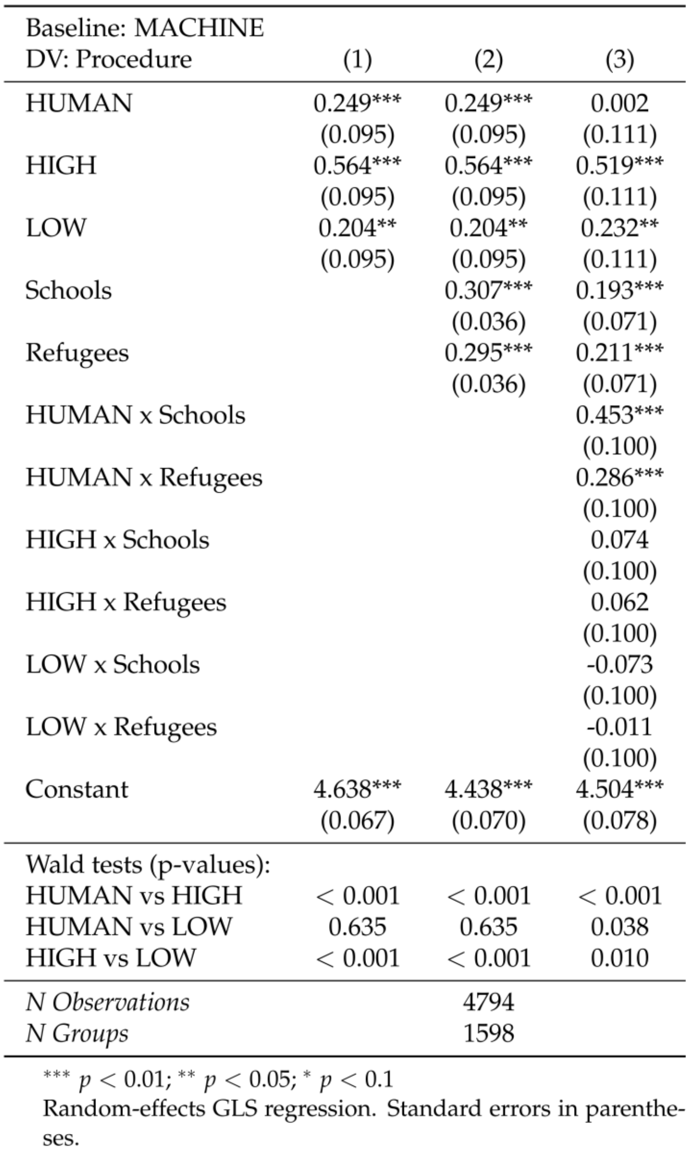
Overall		Scenario-specific	
DV: Procedure		DV: Procedure	
Seq2	-0.252***	Seq2	-0.191**
	(0.071)		(0.075)
Seq3	-0.327***	Seq3	-0.135*
	(0.071)		(0.073)
HIGH	0.333***	Schools	0.359***
	(0.111)		(0.074)
LOW	-0.120	Refugees	0.369***
	(0.111)		(0.073)
MACHINE	-0.345***	Seq2 × Schools	-0.013
	(0.111)		(0.112)
Seq2 × HIGH	-0.035	Seq2 × Refugees	-0.017
	(0.101)		(0.112)
Seq2 × LOW	0.110	Seq3 × Schools	-0.144
	(0.101)		(0.112)
Seq2 × MACHINE	0.137	Seq3 × Refugees	-0.219*
	(0.101)		(0.112)
Seq3 × HIGH	-0.018		
	(0.101)		
Seq3 × LOW	0.115		
	(0.101)		
Seq3 × MACHINE	0.154		
	(0.101)		
Constant	5.080***	Constant	4.801***
	(0.078)		(0.059)
Wald tests (p-values):		Wald tests (p-values):	
Seq2 vs Seq3	0.294	Seq2 vs Seq3	0.450
Seq2 × HIGH vs Seq3 × HIGH	0.867	Seq2 × Schools vs Seq3 × Schools	0.243
Seq2 × LOW vs Seq3 × LOW	0.961	Seq2 × Refugees vs Seq3 × Refugees	0.072
Seq2 × MACHINE vs Seq3 × MACHINE	0.865		
Seq2 × HIGH vs Seq2 × LOW	0.151		
Seq2 × HIGH vs Seq2 × MACHINE	0.089		
Seq2 × LOW vs Seq2 × MACHINE	0.790		
Seq3 × HIGH vs Seq3 × LOW	0.188		
Seq3 × HIGH vs Seq3 × MACHINE	0.089		
Seq3 × LOW vs Seq3 × MACHINE	0.699		
N Observations	4794	*N Observations*	4794
N Groups	1598	*N Groups*	1598

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$
Random-effects GLS regression. Standard errors in parentheses.

C. Additional Analyses of Scenario-Specific Effects

Table 5 presents an analysis of treatment effects on perceived procedural fairness for each scenario with a treatment-scenario interaction in Model 3.

Table 5: Procedural fairness with treatment-scenario interaction

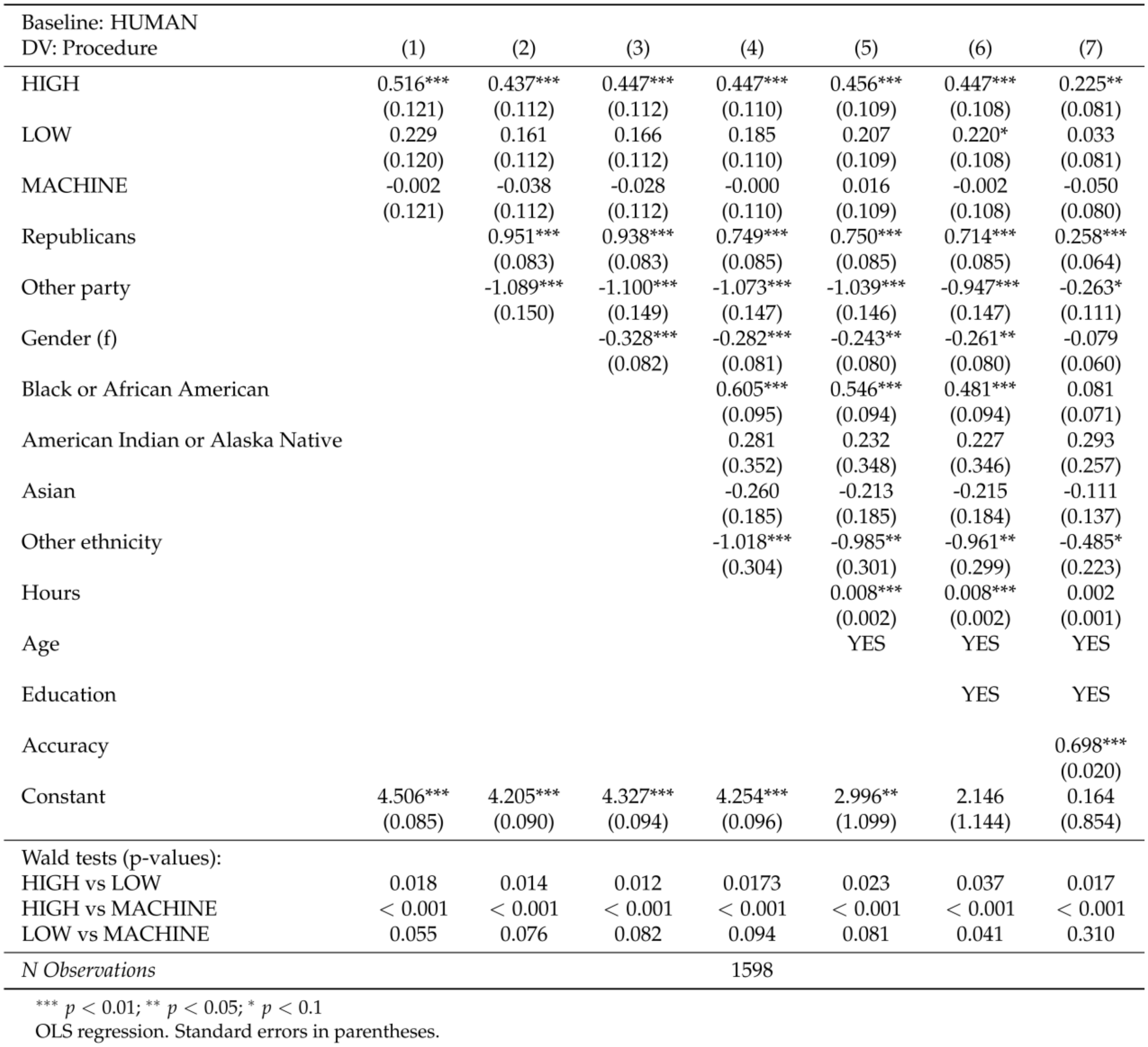


Baseline: MACHINE DV: Procedure	(1)	(2)	(3)
HUMAN	0.249***	0.249***	0.002
	(0.095)	(0.095)	(0.111)
HIGH	0.564***	0.564***	0.519***
	(0.095)	(0.095)	(0.111)
LOW	0.204**	0.204**	0.232**
	(0.095)	(0.095)	(0.111)
Schools		0.307***	0.193***
		(0.036)	(0.071)
Refugees		0.295***	0.211***
		(0.036)	(0.071)
HUMAN x Schools			0.453***
			(0.100)
HUMAN x Refugees			0.286***
			(0.100)
HIGH x Schools			0.074
			(0.100)
HIGH x Refugees			0.062
			(0.100)
LOW x Schools			-0.073
			(0.100)
LOW x Refugees			-0.011
			(0.100)
Constant	4.638***	4.438***	4.504***
	(0.067)	(0.070)	(0.078)
Wald tests (p-values):			
HUMAN vs HIGH	< 0.001	< 0.001	< 0.001
HUMAN vs LOW	0.635	0.635	0.038
HIGH vs LOW	< 0.001	< 0.001	0.010
N Observations		4794	
N Groups		1598	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$
Random-effects GLS regression. Standard errors in parentheses.

Tables 6, 7 and 8 present an analysis of treatment effects on perceived procedural fairness in each scenario used in our experiment (predictive policing, school admissions, refugee-matching).

Table 6: Procedural fairness in the predictive-policing scenario



Baseline: HUMAN							
DV: Procedure	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIGH	0.516***	0.437***	0.447***	0.447***	0.456***	0.447***	0.225**
	(0.121)	(0.112)	(0.112)	(0.110)	(0.109)	(0.108)	(0.081)
LOW	0.229	0.161	0.166	0.185	0.207	0.220*	0.033
	(0.120)	(0.112)	(0.112)	(0.110)	(0.109)	(0.108)	(0.081)
MACHINE	-0.002	-0.038	-0.028	-0.000	0.016	-0.002	-0.050
	(0.121)	(0.112)	(0.112)	(0.110)	(0.109)	(0.108)	(0.080)
Republicans		0.951***	0.938***	0.749***	0.750***	0.714***	0.258***
		(0.083)	(0.083)	(0.085)	(0.085)	(0.085)	(0.064)
Other party		-1.089***	-1.100***	-1.073***	-1.039***	-0.947***	-0.263*
		(0.150)	(0.149)	(0.147)	(0.146)	(0.147)	(0.111)
Gender (f)			-0.328***	-0.282***	-0.243**	-0.261**	-0.079
			(0.082)	(0.081)	(0.080)	(0.080)	(0.060)
Black or African American				0.605***	0.546***	0.481***	0.081
				(0.095)	(0.094)	(0.094)	(0.071)
American Indian or Alaska Native				0.281	0.232	0.227	0.293
				(0.352)	(0.348)	(0.346)	(0.257)
Asian				-0.260	-0.213	-0.215	-0.111
				(0.185)	(0.185)	(0.184)	(0.137)
Other ethnicity				-1.018***	-0.985**	-0.961**	-0.485*
				(0.304)	(0.301)	(0.299)	(0.223)
Hours					0.008***	0.008***	0.002
					(0.002)	(0.002)	(0.001)
Age					YES	YES	YES
Education						YES	YES
Accuracy							0.698***
							(0.020)
Constant	4.506***	4.205***	4.327***	4.254***	2.996**	2.146	0.164
	(0.085)	(0.090)	(0.094)	(0.096)	(1.099)	(1.144)	(0.854)
Wald tests (p-values):							
HIGH vs LOW	0.018	0.014	0.012	0.0173	0.023	0.037	0.017
HIGH vs MACHINE	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
LOW vs MACHINE	0.055	0.076	0.082	0.094	0.081	0.041	0.310
N Observations				1598			

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$
OLS regression. Standard errors in parentheses.

Table 7: Procedural fairness in the school-admissions scenario

Baseline: HUMAN							
DV: Procedure	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIGH	0.138	0.085	0.093	0.091	0.091	0.086	-0.097
	(0.107)	(0.103)	(0.103)	(0.102)	(0.102)	(0.101)	(0.081)
LOW	-0.297**	-0.339**	-0.335**	-0.320**	-0.317**	-0.297**	-0.347***
	(0.107)	(0.103)	(0.103)	(0.102)	(0.102)	(0.101)	(0.081)
MACHINE	-0.455***	-0.476***	-0.468***	-0.448***	-0.444***	-0.459***	-0.386***
	(0.107)	(0.103)	(0.103)	(0.102)	(0.101)	(0.101)	(0.080)
Republicans		0.660***	0.650***	0.499***	0.498***	0.467***	0.193**
		(0.076)	(0.076)	(0.079)	(0.079)	(0.079)	(0.064)
Other party		-0.581***	-0.590***	-0.571***	-0.542***	-0.458***	-0.159
		(0.138)	(0.138)	(0.136)	(0.136)	(0.137)	(0.110)
Gender (f)			-0.263***	-0.235**	-0.219**	-0.228**	-0.088
			(0.075)	(0.075)	(0.075)	(0.075)	(0.060)
Black or African American				0.440***	0.392***	0.333***	0.055
				(0.088)	(0.088)	(0.088)	(0.071)
American Indian or Alaska Native				-0.053	-0.078	-0.089	-0.249
				(0.326)	(0.324)	(0.323)	(0.258)
Asian				-0.555**	-0.531**	-0.565**	-0.220
				(0.171)	(0.173)	(0.172)	(0.138)
Other ethnicity				-0.484	-0.468	-0.469	-0.295
				(0.281)	(0.280)	(0.279)	(0.223)
Hours					0.007***	0.007***	0.002
					(0.002)	(0.002)	(0.001)
Age					YES	YES	YES
Education						YES	YES
Accuracy							0.648***
							(0.022)
Constant	5.152***	4.925***	5.023***	4.996***	5.737***	5.857***	2.951***
	(0.076)	(0.083)	(0.087)	(0.089)	(1.023)	(1.069)	(0.858)
Wald tests (p-values):							
HIGH vs LOW	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.002
HIGH vs MACHINE	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
LOW vs MACHINE	0.138	0.183	0.195	0.209	0.210	0.110	0.633
N Observations				1598			

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$
OLS regression. Standard errors in parentheses.

Table 8: Procedural fairness in the refugee-matching scenario

Baseline: HUMAN							
DV: Procedure	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIGH	0.292**	0.234*	0.240*	0.239*	0.230*	0.227*	0.136
	(0.104)	(0.100)	(0.100)	(0.099)	(0.098)	(0.097)	(0.074)
LOW	-0.067	-0.112	-0.109	-0.102	-0.108	-0.096	-0.045
	(0.104)	(0.099)	(0.099)	(0.099)	(0.098)	(0.098)	(0.075)
MACHINE	-0.288**	-0.310**	-0.304**	-0.286**	-0.280**	-0.300**	-0.190*
	(0.104)	(0.100)	(0.099)	(0.099)	(0.098)	(0.097)	(0.074)
Republicans		0.755***	0.748***	0.618***	0.617***	0.583***	0.280***
		(0.074)	(0.074)	(0.076)	(0.077)	(0.077)	(0.059)
Other party		-0.550***	-0.556***	-0.540***	-0.501***	-0.459***	-0.149
		(0.133)	(0.133)	(0.132)	(0.131)	(0.132)	(0.101)
Gender (f)			-0.176*	-0.144*	-0.112	-0.119	0.015
			(0.073)	(0.072)	(0.072)	(0.072)	(0.055)
Black or African American				0.413***	0.359***	0.315***	0.000
				(0.085)	(0.085)	(0.085)	(0.066)
American Indian or Alaska Native				0.461	0.431	0.443	0.160
				(0.315)	(0.313)	(0.312)	(0.238)
Asian				-0.135	-0.032	-0.046	0.103
				(0.166)	(0.166)	(0.166)	(0.127)
Other ethnicity				-0.760**	-0.749**	-0.733**	-0.444*
				(0.272)	(0.270)	(0.269)	(0.205)
Hours					0.005**	0.005**	0.000
					(0.002)	(0.002)	(0.001)
Age					YES	YES	YES
Education						YES	YES
Accuracy							0.675***
							(0.020)
Constant	5.002***	4.732***	4.797***	4.745***	4.512***	4.790***	1.973*
	(0.074)	(0.080)	(0.084)	(0.086)	(0.987)	(1.031)	(0.790)
Wald tests (p-values):							
HIGH vs LOW	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.001	0.015
HIGH vs MACHINE	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
LOW vs MACHINE	0.034	0.047	0.050	0.063	0.079	0.037	0.053
N Observations				1598			

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$
OLS regression. Standard errors in parentheses.

D. Instructions

Police

Decision making procedure ***(HUMAN)***

The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.

Based on his or her risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.

Decision making procedure ***(HIGH)***

The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer.

The decision will never be based on the computer algorithm alone. The police officer will always conduct his or her own analysis, that means, the police officer will in each case conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.

Based on the risk assessment of the computer algorithm and his or her own risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.

Decision making procedure ***(LOW)***

The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer.

The decision will usually be based on the computer algorithm alone. The police officer will sometimes conduct his or her own analysis, that means, the police officer will in some cases conduct an in-depth analysis of the case material, and assess the risk of violent assaults in certain areas of the city.

Based on the risk assessment of the computer algorithm and - only if conducted - his or her own risk assessment, the police officer will order or not order bodily searches in a certain area of the city. The police officer has discretion in this decision.

Decision making procedure ***(MACHINE)***

The senior police officer in charge will collect information on previous cases of violent assaults in the city. Then the police officer will use a computer algorithm to assess the risks of violent assaults in certain areas of the city. The computer algorithm will conduct an in-depth analysis of the case material and present its risk assessment to the police officer.

The decision will be based on the computer algorithm’s assessment alone.

Based on its risk assessment, the computer algorithm will order or not order bodily searches in a certain area of the city. The police officer cannot override the decision of the computer algorithm and has no discretion in this decision.

How fair do you consider the procedure by which the police come to their decision? (1=very unfair, 7=very fair)

How accurately do you think the police will assess the risk of violent crimes in the city? (1=not accurately at all, 7=extremely accurately)

To what extent would you want the decision making procedure to be used in a case if you were personally concerned? (1=not at all, 7=to a large extent)

To what extent would you want the decision making procedure to be used in a case for the general public? (1=not at all, 7=to a large extent)

Schools

Decision making procedure ***(HUMAN)***

The school admissions board will conduct an interview with the applicant and collect additional information on the applicant’s school history, extracurricular engagement, the distance between the school and the home, and family support. Then the admissions board will conduct an in-depth analysis of the case material and assess the applicant’s success probability.

Based on its assessment of the success probability, the admissions board will accept or reject the applicant. The admissions board has discretion in this decision.

Decision making procedure ***(HIGH)***

The school admissions board will conduct an interview with the applicant and collect additional information on the applicant’s school history, extracurricular engagement, the distance between the school and the home, and family support. Then the admissions board will use a computer algorithm to assess the applicant’s success probability. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the applicant's success probability to the admissions board.

The decision will never be based on the computer algorithm alone. The admissions board will always conduct its own analysis, that means, the admissions board will in each case conduct an in-depth analysis of the case material, and assess the applicant's success probability.

Based on the assessment of the success probability of the computer algorithm and its own assessment of the success probability, the admissions board will accept or reject the applicant. The admissions board has discretion in this decision.

Decision making procedure ***(LOW)***

The school admissions board will conduct an interview with the applicant and collect additional information on the applicant's school history, extracurricular engagement, the distance between the school and the home, and family support. Then the admissions board will use a computer algorithm to assess the applicant's success probability. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the applicant's success probability to the admissions board.

The decision will usually be based on the computer algorithm alone. The admissions board will sometimes conduct its own analysis, that means, the admissions board will in some cases conduct an in-depth analysis of the case material and assess the applicant's success probability.

Based on the assessment of the success probability of the computer algorithm and - only if conducted - its own assessment of the success probability, the admissions board will accept or reject the applicant. The admissions board has discretion in this decision.

Decision making procedure ***(MACHINE)***

The school admissions board will conduct an interview with the applicant and collect additional information on the applicant's school history, extracurricular engagement, the distance between the school and the home, and family support. Then the admissions board will use a computer algorithm to assess the applicant's success probability. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the applicant's success probability to the admissions board.

The decision will be based on the computer algorithm's assessment alone.

Based on its assessment of the success probability, the computer algorithm will accept or reject the applicant. The admissions board cannot override the decision of the computer algorithm and has no discretion in this decision.

How fair do you consider the procedure by which the school admissions board comes to its decision? (1=very unfair, 7=very fair)

How accurately do you think the school admissions board will assess the probability that the applicant will eventually graduate? (1=not accurately at all, 7=extremely accurately)

To what extent would you want the decision making procedure to be used in a case if you were personally concerned? (1=not at all, 7=to a large extent)

To what extent would you want the decision making procedure to be used in a case for the general public? (1=not at all, 7=to a large extent)

Refugees

Decision making procedure ***(HUMAN)***

The case manager will conduct an interview with the refugee and collect information on the refugee’s origin, education, linguistic skills, and family status. Then the case manager will conduct an in-depth analysis of the case material and assess the probability of successful employment.

Based on his or her assessment of the probability of successful employment, the case manager will assign the refugee to a certain location. The case manager has discretion in this decision.

Decision making procedure ***(HIGH)***

The case manager will conduct an interview with the refugee and collect information on the refugee’s origin, education, linguistic skills, and family status. Then the case manager will use a computer algorithm to assess the probability of successful employment. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the probability of successful employment to the case manager.

The decision will never be based on the computer algorithm alone. The case manager will always conduct his or her own analysis, that means, the case manager will in each case conduct an in-depth analysis of the case material and assess the probability of successful employment.

Based on the assessment of the probability of successful employment of the computer algorithm and his or her own assessment of the probability of successful employment, the case manager will assign the refugee to a certain location. The case manager has discretion in this decision.

Decision making procedure ***(LOW)***

The case manager will conduct an interview with the refugee and collect information on the refugee’s origin, education, linguistic skills, and family status. Then the case manager will use a computer algorithm to assess the probability of successful employment. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the probability of successful employment to the case manager.

The decision will usually be based on the computer algorithm alone. The case manager will sometimes conduct his or her own analysis, that means, the case manager will in some cases conduct an in-depth analysis of the case material and assess the probability of successful employment.

Based on the assessment of the probability of successful employment of the computer algorithm and - only if conducted - his or her own assessment of the probability of successful employment, the case manager will assign the refugee to a certain location. The case manager has discretion in this decision.

Decision making procedure ***(MACHINE)***

The case manager will conduct an interview with the refugee and collect information on the refugee’s origin, education, linguistic skills, and family status. Then the case manager will use a computer algorithm to assess the probability of successful employment. The computer algorithm will conduct an in-depth analysis of the case material and present its assessment of the probability of successful employment to the case manager.

The decision will be based on the computer algorithm’s assessment alone.

Based on its assessment of the probability of successful employment, the computer algorithm will assign the refugee to a certain location. The case manager cannot override the decision of the computer algorithm and has no discretion in this decision.

How fair do you consider the procedure by which the immigration authority comes to its decision? (1=very unfair, 7=very fair)

How accurately do you think the immigration authority will assess the probability that the refugee will successfully find employment? (1=not accurately at all, 7=extremely accurately)

To what extent would you want the decision making procedure to be used in a case if you were personally concerned? (1=not at all, 7=to a large extent)

To what extent would you want the decision making procedure to be used in a case for the general public? (1=not at all, 7=to a large extent)

1. You have seen three different scenarios. Please rank these scenarios according to the severeness of the decision for the recipient from 1 (least severe) to 3 (most severe).

2. In this survey, you have been asked to assess the fairness of several decision making procedures by public officials. Please state shortly for what reasons you decided the way you did, especially on which criteria you based your evaluation of the fairness of the procedure (keywords are sufficient).

3. How old are you?

4. What is your highest educational degree?

5. What is your gender?

6. What is your ethnicity?

7. Which political party do you feel closest to?

8. How many hours per week do you spend online doing tasks for money?